Biometrics in Plant Breeding: Applications of Molecular Markers

Proceedings of the Ninth Meeting of the EUCARPIA Section Biometrics in Plant Breeding, 6 – 8 July 1994, Wageningen, the Netherlands

Edited by

J.W. van Ooijen & J. Jansen

DLO-Centre for Plant Breeding and Reproduction Research (CPRO-DLO), Wageningen, the Netherlands

EUCARPIA

European Association for Research on Plant Breeding
Association Européenne pour l'Amélioration des Plantes
Europäische Gesellschaft für Züchtungsforschung

Ninth Meeting of the Eucarpia Section Biometrics in Plant Breeding

Organizing Committee

I. Bos (Agricultural University, Wageningen), chairman
J. Jansen (CPRO-DLO, Wageningen), secretary
P. Stam (CPRO-DLO/Agricultural University, Wageningen), treasurer
L. Beerepoot (Barenbrug Holland, Oosterhout)
M.T. Morales (VanderHave, Rilland)
H.R.M. Kentie, H. Bloemhof (IAC, Wageningen)

How to order this book:

This book can be ordered by sending a request to: CPRO-DLO, Dept. Pop. Biology, stating the reference: "4408 / Proceedings Biometrics".

By mail: P.O. Box 16
6700 AA Wageningen
the Netherlands

By fax: + 31 8370 18094

CIP DATA KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Biometrics

Biometrics in plant breeding: applications of molecular markers / ed.
by J.W. van Ooijen & J. Jansen. - Wageningen : CPRO-DLO. - Ill.
Proceedings of the Ninth Meeting of the EUCARPIA Section Biometrics
in Plant Breeding, 6–8 July 1994, Wageningen. - With ref.
ISBN 90-73771-13-7
Subject headings: biometrics / molecular markers / plant breeding.

Preface

This book contains the proceedings of the Ninth Meeting of the Eucarpia Section Biometrics in Plant Breeding. The main theme of this meeting was *Applications of Molecular Markers.* The papers on this theme reflect to a large extent the current status of biometrical research on applications of molecular markers in plant breeding and variety registration. They demonstrate that practical applications of molecular markers require intensive collaboration between biometricians, plant breeders and molecular biologists, now and in the future. The papers on more classical issues, like phenotypic stability and design of variety trials, show that a continuing biometrical input is needed to improve the efficiency of breeding systems and variety testing.

Acknowledgements

The organizing committee kindly appreciates generous support from the following companies and institutions:

Agrico Research
Barenbrug Holland BV
Bruinsma Seeds BV
Cebeco Zaden BV
De Ruiter Zonen CV
Department of Plant Breeding, Wageningen Agricultural University
DLO-Centre for Plant Breeding and Reproduction Research
EUCARPIA
KeyGene NV
KWS Kleinwanzlebener Saatzucht AG
LEB-fonds
Sluis Vegetables
Stichting Broekemafonds
VanderHave
Zeneca Seeds

Contents

Opening address

Invited papers

Papers

Posters

Opening address

K.B. Geling, Royal Vanderhave Group, P.O. Box 1, 4420 AA Kapelle, The Netherlands

Mr. Chairman, Ladies and Gentlemen,

Thanks to the Dutch Organizing Committee for the invitation to present the Opening Address to your conference.

First of all I wish you a most cordial welcome to Holland, not only on request of the Organizing Committee, but also on behalf of the Dutch Seed Industry. The Seed Industry favours also to stimulate that beautiful ideas may arise during these 3 days of the meeting of this section here in Wageningen.

The Dutch Seed Industry itself is not in a very flourishing economic situation nowadays, this mainly as a consequence of the depressed economy in the domain of agriculture in general. This economic situation forces the industry to realize changes and reconsider its strategy. Also, research efforts will not escape from this consideration.

Since plant breeding is a relatively slow process, the Seed Industry is economically very interested in decreasing the amount of time involved in such a process. It therefore shows a high priority in any possibility to reduce time and costs invested in breeding. The main topic of this meeting: "The application of molecular markers" might develop some good potentials to reach the goal of breeding still better varieties, but at a more acceptable cost level.

Although I realize that this goal is not to be reached by tomorrow, we can, however, conclude that the recent technical progress made in plant biotechnology makes us confident about the positive contribution for the development of future varieties. This progress has been achieved on two important fields:

- the development of molecular markers leading to marker-assisted breeding;
- the introduction of alien genes leading to transgenic varieties.

The last mentioned field has its own specialties and difficulties, of which momentarily public acceptance, and therefore being considered as Novel Food, is the most critical hurdle to pass. However, this is an item not covered by this meeting and therefore needs no digression at the moment.

The development of molecular markers has hardly any relevance to the acceptance by the public, so you do not have to deal with this major hurdle. A prerequisite for the development of markers to be used in applied plant breeding is, however, that the costs of this indirect selection technique are ultimately not higher than those of the current direct selection. This because of the fact that, in contrary to transgenic breeding, the marker-assisted breeding will not cause an additional added value of the final variety on the market. So, the positive effects cannot be translated in selling these varieties against a higher price, but should be found in a lower cost price through time-saving or through lower real costs by replacing current expensive testing methods. The possibility to identify good combining parents to reach an optimal heterosis in hybrids is a good example of an enormous potential for cost savings. Since one of the lectures deals with the efficiency of marker-assisted selection I trust that this has got already your attention.

Another item in the context of this meeting I will shortly refer to, is a recent discussion in Assinsel. The UPOV-convention was in 1991 revised and granted also certain rights for the so-called "essentially derived varieties" to the holder of the rights on the original variety. Therefore, we are now faced with the question to define the degree of conformity between varieties. Your skill in marker-development can be of great help in establishing a clear definition for that purpose.

In a position paper Assinsel stated that in the interpretation of the UPOV-legislation the two concepts of distinctness and conformity of varieties should not be confused. For the time being they imply that different tools will be used for defining the two concepts. This means:

a. distinctness should be assessed with the help of morphological and physiological characters, which for practical reasons should be easily to identify, *e.g.*, in field certification. So distinctness is of phenotypic nature. Biochemical markers can only add precision if considering distinctness;
b. conformity or essential derivation is particularly a genotypical question and therefore DNA-analysis is a useful tool to prove conformity between varieties. This analysis should, however, not be used to demonstrate the uniformity of a variety within the framework of DUS-legislation. Such an analysis would increase the costs to check the uniformity of the variety in maintenance breeding to an unacceptable level.

Depending on the evolution of molecular markers, this position can be changed in the future, although now it is likely to be a challenge for the marker-assisted selection technology.

Mr. Chairman, with these few words I declare the Ninth Meeting of the Eucarpia Section Biometrics in Plant Breeding for opened and I wish you a very flourishing and fruitful congress.

Recent technical developments for possible use in registration of new plant varieties

Michael S. Camlin & Trevor J. Gilliland, Department of Agriculture for Northern Ireland, Plant testing Station, Crossnacreevy, Belfast BT6 9SH, U.K.

Key words
registration, varieties, protein electrophoresis, DNA fingerprinting

Abstract

Through protein electrophoresis and, more recently, DNA fingerprinting techniques, rapid advances are taking place in genetic studies across all the biological sciences.

Within the plant sciences and plant breeding in particular the possibility of developing rapid and simple methodologies for varietal identification and description is now provided. The identification of varieties already registered and protected is however rather different from the registration of new varieties under Plant Breeders' Rights legislation and there are important differences in philosophy to be considered.

In a rapidly changing situation with progress in methodology continuing at a rapid pace, this paper attempts to identify the principles which may be applicable and to examine how the most recent and possible future technical advances may best be exploited within the context of plant variety protection and registration.

The views expressed in this paper are personal opinions and the paper does not represent any statement of policy on behalf of either the UK Testing Authorities or of UPOV.

Introduction - A future scenario

If the promised contribution which molecular biology is said to be able to make to plant breeding bears full fruit then we should in future begin to see a rather more complex situation taking shape with respect to plant variety and seed marketing.

A farmer, a few years from now, when advance ordering seed for autumn sowing of his winter wheat crop may see an advertisement for the new variety Millennium. He might be very interested in this new variety because he may have heard that it has the same yield potential and field characteristics as the variety Centurion which he has previously been growing but that it has better resistance to *Septoria* and mildew provided by new gene constructs from two agrochemical companies and increased resistance to grass weed herbicides from an American multinational.

For the farmer, the purchasing of the seed of the variety he requires is exactly as it is today, but behind the scenes a much more complex situation has arisen. The intellectual property protection provided for the variety as a whole, the UPOV system of Plant Breeders' Rights, has been joined by a further degree of intellectual property protection provided by the patent system for the various genes which have been incorporated into the variety. This in itself poses no insurmountable problem and is not so different from the situation which currently exists in the automotive or computer industries where products have within them a collection of patents and licencing agreements for various components about which the producer knows and cares little.

The more difficult question however, is how, with plant varieties, we can move from the current system of examination for protection based upon the appearance, morphology or phenotype of the variety to the types of systems which will be required in this developing situation for adequate protection of the existing variety Centurion, registration of the new variety entity Millennium and also protection of the genetic 'components' within it.

The UPOV convention

The new 1991 UPOV Convention which was the subject of much discussion a few years ago has provided for this sort of situation. For the first time the concept of Essential Derivation allows for the possibility of fair recompense in terms of intellectual property protection for the holder of Plant Breeders' Rights of the initial variety, for the breeder of the new improved variety and also for the patent holder of any gene constructs used to 'improve' it.

The challenge which now has to be faced on the technical side is how to harness the new technologies to allow clear identification of the variety and fully meet the conditions for its protection set out in Article 5 of the UPOV Convention. Showing novelty is easy as the variety simply has to be basically a new entity, but the question of Distinctness will become much more complex and is bound up with the question of definition of the variety itself. Also the questions of Uniformity and Stability, taken together, need to be examined further within the new situation.

Distinctness

On Distinctness the 1991 UPOV convention states at Article 7 that "The *variety* shall be deemed to be distinct if it is *clearly distinguishable* from any other variety whose existence is a matter of common knowledge ..."

The former 1978 Convention at Article 6(1)(a) made the proviso that the characteristics which would permit a variety to be defined and distinguished must be capable of precise recognition and description. The new 1991 Convention makes no such statement and thus appears to leave the way clear for more complex multivariate or genetic distance measurements using a range of different characteristics.

Variety

The statement on Distinctness in Article 7 of the 1991 Convention has to be taken together with the definition of variety in Article 1 where the variety is said to be:

(i) 'defined by the *expression of the characteristics* resulting from a given *genotype* or combination of genotypes'

(ii) 'distinguished from any other plant grouping by the *expression* of at least one of the said characteristics'

(iii) 'considered as a unit with regard to its suitability for being propagated unchanged'.

The third condition effectively covers Uniformity and Stability. From the other two conditions the variety can be firstly defined and secondly distinguished by the expression of a characteristic or characteristics resulting from a given genotype. This also appears to allow for multivariate or genetic distance measurements, but most significantly clearly requires the expression of genotypic differences through to the phenotype.

Essential derivation

Finally the matter of "clearly distinguishable" within the definition of Distinctness in Article 7 is tempered by the consideration of Essential Derivation at Article 14 where the scope of the breeders' right is considered in more detail.

From the very detailed Article 14, the key statements with respect to definition of Essential Derivation are at clause 5, sub-clauses b(i), (ii) and (iii) which read as follows:

"... a *variety* shall be deemed to be essentially derived from another *variety* ("the initial variety") when

(i) it is predominately derived from the initial *variety*, ... while retaining the *expression of the essential characteristics* that result from the *genotype* or combination of genotypes of the initial *variety*.

(ii) it is *clearly distinguishable* from the initial *variety* and

(iii) except for the differences which result from the act of derivation, it conforms to the initial *variety* in the *expression of the essential characteristics* that result from the *genotype* or combination of genotypes of the initial variety.

Here, once more, the expression of genotypic differences through to the phenotype is clearly required for any consideration of the variety concept.

These three Articles together confirm the intention in the 1991 UPOV Convention to define three types of varieties;

(1) the Clearly Distinct and non-derived variety.

(2) the Clearly Distinct and essentially derived variety.

(3) the Non-distinct variety.

The first two types of varieties are protectable within the terms of the Convention but the third is not.

The concept of Essential Derivation in the new Convention was aimed at providing better protection for conventional plant breeders with protected varieties in the face of a possible onslaught from genetic manipulation by molecular biologists. It was also thought that the concept might ease the dilemma for testing authorities posed by minimum (genetic) distance, the relevance of which had long been debated with respect to the establishment of Distinctness. It is however important to note that while UPOV introduces the concept of Essential Derivation, it confines itself to direct concern only with Distinctness for the award of Plant Breeders' Rights. Any consideration of Essential Derivation falls outside this direct concern and is left to the industry to resolve. How large a role the registration authorities should play in advising on Essential Derivation is still under discussion.

In considering the technical aspects of examining varieties for their eligibility for the award of Plant Breeders' Rights there are several key words and phrases across these three Articles which should be briefly examined: 'genotype' (and phenotype), 'expression of characteristics', 'clearly distinguishable' and 'variety'.

Perhaps these can best be summarised in Figure 1 which is an interpretation of the relationship between genome, genotype and phenotype as relevant to variety protection. Diagrammatically represented is the sequence of basic building blocks from simple nucleotide base-pairs through to the complexity of whole-plant organisation, together with the descriptive terms genome, genotype and phenotype.

On this basis, the UPOV definition of variety which is linked in Article I to "the

Figure 1. Interpretation of genome, genotype and phenotype as relevant to variety protection

Chromosomes Nucleotide sequences	Genome
⇓ ⇓	
Genes (expressed regions)	Genotype
⇓ ⇓	
Proteins Enzymes Biostructures, Biocompounds Plant Morphology Physiology	Phenotype

expression of the characteristics resulting from a given genotype" comes in only after the level of organisation of DNA into meaningful genes. Distinctness, which in Article 7 only requires one variety to be "clearly distinguishable" from another is nevertheless also linked to the variety concept as defined in Article 1. Distinctness can thus only be shown between varieties using characteristics known to derive from the expressed parts of the genome. Such areas constitute the genotype which determines the phenotype either as measurable morphological or physiological traits or as hidden, possibly polygenic, contributors to the overall phenotype, for example in characteristics such as plant height or intermediate metabolites.

This interpretation, if accepted, has implications for the way in which the new technologies might be applied. This is especially relevant to some of the more basic DNA fingerprinting techniques which consider only genomic organisational and structural differences and do not necessarily consider their expressed phenotypic effects.

Recent technical developments

With the development of protein electrophoresis and, more recently, DNA fingerprinting techniques, rapid advances have taken place in genetic studies across all the biological sciences. Within the plant sciences and plant breeding in particular the possibility now exists for developing and standardising rapid and simple methodologies for varietal identification and recognition. A wide range of techniques and methods such as DNA profiling and gene probing using RFLP and the various PCR based techniques including RAPD's will provide routine tests for the identification of differences between varieties at the genomic level. There is now a large number of papers covering the use of the techniques for various purposes including gene mapping and the study of evolutionary relationships across a wide range of crops.

Most interestingly, an increasing number of papers is now specifically addressing the problem of identification of varieties or lines using several different techniques and systems for example: RFLP with Barley - Bunce *et al.* (1986); RAPD with Brassicas - Hu & Quiros (1991); RFLP with Maize - Smith *et al.* (1991); RAPD with Maize - Welsh *et al.* (1991); RAPD with Potato - Demeke *et al.* (1993); RAPD with Wheat - Joshi & Nguyen (1993) and RFLP and RAPD with Tomato - Williams & St. Clair (1993).

From the point of view of variety registration, the potential ease with which varieties can be identified using DNA profiling and gene probing techniques in the laboratory is undoubtedly a great attraction for breeders and testing authorities alike. This is particularly so as only a few hours of work is required in contrast to several months or even several growing seasons in the field using conventional morphological or physiological characteristics. Although the costs of individual identifications remain high they are probably no greater than those involved in a full morphological examination and, most significantly for certain crops, can provide a stable and more environmentally independent identity for the variety.

To many involved in variety registration these attributes will seem very similar to those reviewed and considered some 10 or more years ago for protein electrophoresis (Wright, Gilliland & Camlin 1983), yet it has taken a considerable time to reach the point where the use of electrophoresis in variety registration is now becoming accepted by breeders and testing authorities alike.

Accepting the principles outlined in the UPOV convention it is interesting to consider some of the methods becoming available together with conventional morphological characteristics.

Morphological observations

There is no doubt that in the past these characteristics have been closest to the way in which the farmer considers varietal identity. However, this may not still be the case in the future with the inclusion in varieties of the new genetic components outlined earlier. The farmer may wish to be sure that certain genes conferring useful agronomic attributes are in the variety and may also begin to identify with these.

Part of the problem now facing registration authorities with morphological characteristics is the history of having in the past granted protection to varieties on the basis of several different principles. The word 'important' with respect to Distinctness in Article 6 of the 1978 UPOV Convention whose definition caused much discussion has now been dropped and replaced with 'clearly distinguishable' in Article 7 of the 1991 Convention. However, problems still remain because of the previous interpretation of 'important' in two different senses. Substantial polygenic differences across major tracts of the genome and small single gene differences have been given equivalent status in the granting of PBR protection. For example, on the one hand basic and polygenic features such as plant height have been used for Distinctness for a range of crops. On the other hand there has also been widespread application of single gene characteristics. Some have been commercially important as in the resistance to *Bremia* in lettuce or variation in flower colour in various ornamentals while others have been important only for Distinctness, for example, phenol reaction in wheat or DDT reaction in barley.

Thus, while Distinctness can still easily be determined with the morphological characteristics, it will now be very difficult indeed to sort out the problems of Essential Derivation and genetic distance using such characteristics alone. It is perhaps in this area that the newer technologies may be of considerable potential in the future.

Protein typing

With the various electrophoretic methods available for examining plant proteins or enzymes the situation is not very far removed from that with conventional morphological characteristics. The question of expression is not an issue as these proteins and enzymes are clearly part of the phenotype. UPOV, after protracted discussions across many years has now also established the principle of use of protein electrophoresis only when the genetic basis is understood. This is a useful firm foundation and is allowing progress with application of several alternative methods to the examination of variety Distinctness in a range of crops which have unique problems.

In this context, the recent adoption of standard methods by UPOV for glutenins in wheat and hordeins in barley is particularly significant in clearly linking the application of electrophoretic examination of these seed proteins to an understanding of the genetic basis for the differences obtained on gels. The acceptance of electrophoretic methods in wheat and barley thus involves the adoption by UPOV of a very important basic principle which extends the philosophy in Article 1 concerning "expression of the characteristics resulting from a given genotype". Thus gel patterns or 'barcodes' may be perfectly satisfactory for identification of pre-registered varieties but for the *de novo* registration of new varieties a proper genetic interpretation of the bands is necessary.

This principle of understanding the genetic basis of the band patterns recorded on the gel must be firmly adhered to if muddled thinking is to be avoided between what is required on the one hand for simple identification and on the other for variety registration.

DNA profiling

There are many definitions across this whole area of technology, but in this paper the term DNA profiling is confined to an examination of the organisation and structure of the genome without any interpretative effort being made. Such straightforward profiling methods would appear to have excellent potential for use in variety and plant identification. However, the fact that such methods generally do not identify the presence of genes and simply chop up the genome into fragments to be compared as a method of variety classification means that individual gene expression is not involved and certainly no link to genetic interpretation is provided. Such methods may therefore in future have to be confined to use for identification purposes only and may be inappropriate for variety registration and protection purposes.

Thus taking the philosophy on genetic interpretation which has now been established for electrophoresis together with the concept of expression outlined earlier, it is clear that some of the DNA profiling methodologies which simply examine genomic structure may not be entirely appropriate for plant variety registration. They may however be useful in consideration of Essential Derivation, provided mapping has been carried out to show good distribution across the genome.

Gene probing

Where the genome is being probed for the presence of or differences in recognised genes which have an expression in the phenotype, then a different situation can be considered to exist. With gene probing, the presence or otherwise in a variety of selected genes or differences in their DNA make-up can be determined.

Such possibilities will clearly be most useful in investigation of patent protection for novel genes in different varieties. However, we have also to recognise the possibility of determining Distinctness, perhaps in an Essentially Derived situation on the basis of the addition of a novel gene, fully expressed in the phenotype which may be inserted into a variety.

The gene probing approach could also be used for a valid genetic distance measurement if a set of genes, shown to be well distributed across the genome was examined. However, DNA profiling methods will probably be equally appropriate here and will be more straightforward in application.

A summary of the possible applications for these different methodologies is presented in Table 1.

Table 1. Summary of possible and most appropriate applications of methodologies across intellectual property protection systems in plant varieties

	Variety distinctness and registration	Variety essential derivation	Gene patent protection
Morphological observations	✓	?	?
Protein typing	✓[1]	?	?
DNA profiling	x	✓[2]	x
Gene probing	✓[3]	?	✓

[1] Assuming genetic interpretation of gel band patterns is known.

[2] Assuming good distribution of markers across genome.

[3] Assuming expression in the phenotype is shown and is uniform and stable within the variety

Identification and registration

There are clearly common considerations across electrophoresis, DNA profiling and gene probing with respect to the difference between their application to the more straightforward use for identification of already-registered varieties or their use for registration purposes and granting of PBR protection to new varieties. To move from one to the other may seem to some to be an entirely logical step but there are several additional factors with regard to variety registration which are not implicated in simple variety identification.

The most significant point of all may be the overall effects of reducing of minimum distances between varieties through wholesale use of inappropriate techniques for registration and the knock-on effects this could have upon variety protection. This is an issue which has already stimulated considerable debate between breeders and registration authorities and has yet to be fully resolved.

These are however quite basic differences in philosophy involved in identification and registration. Identification is much more straightforward and simply involves the determination of the identity of an unknown but already protected variety sample by comparison with an established reference collection or variety description. Registration, on the other hand, involves the *de novo* establishment of the identity and Distinctness of a reputedly novel variety selection, provision of a definitive description and determination of its eligibility for protection.

Another important aspect within variety registration is the question of Uniformity and associated Stability of varieties. It is clear that morphological examinations for Uniformity will continue to be essential as normally it is not possible to clearly establish firm links between Uniformity in protein or genome composition and plant morphology. This means that the new techniques will have to be used alongside the morphological examinations and of course there are implications here for costs both in registration and in variety maintenance.

The UPOV philosophy of 'last resort' use for electrophoretic characteristics for Distinctness represents a common-sense approach, with the accompanying Uniformity requirements for these more complex and expensive examinations confined to cases where they provide the only means of establishing the Distinctness of a variety. Thus a commercial decision can be taken by the breeder as to whether the market potential of the variety warrants the extra cost of registration by sophisticated methodology and indeed the further extra costs which may be incurred during variety maintenance. A similar common-sense philosophy can be imagined for characteristics determined through gene probing methods.

Practical considerations

These include the obvious and more straightforward points such as the need for standardised methodologies, proof of reproducibility between laboratories and agreed interpretation of information from gels which should be relatively easy to resolve in time. However, there are other practical problems within the various technologies which may be of particular relevance to biometricians.

Traditionally, protein electrophoresis and the various DNA profiling and gene probing methods have been essentially the domain of the laboratory chemist. As such the procedures have been regarded in a similar fashion to, for example, inorganic reactions. Procedural errors were largely considered first and their resolution given priority over possible environmental or biotic variation. Therefore, as in inorganic analyses, sample repeats would be taken to provide proof that the procedure was conducted with complete accuracy and the conclusion drawn that the result could be regarded as an absolute. As these techniques have been increasingly adopted to study an ever more diverse range of biological and particularly botanical problems, shortcomings have become evident and a more considered approach to assessing the total variability encompassed in these systems is now required. Various sources of error must be determined by the biologist and biometrician and these can probably be classified as coming from two separate areas.

Biotic variability: This will involve the normal type of experimental variability encountered in any biological study. To take it into account will require the measurement of sampling errors and accounting for such factors as ontogenetic changes and genetic diversity within the target material (*e.g.*, variety). This would be very dependent on the mode of reproduction of the variety being examined (*i.e.*, allogamous, autogamous, clonal, hybrid etc.) but should present relatively straightforward statistical problems, similar to those already dealt with within UPOV for conventional characteristics.

Procedural variability: This is specific to these biotechnological processes and is a consequence of the sensitivity of the techniques employed. The sources of variability occur due to minute differences in experimental conditions (extraction efficiency, buffer pH, gel quality, voltage, etc.), which can cause small changes in band intensity and band position, and lead to compression or stretching of gel patterns.

The number of individual techniques now being employed is vast and if each had to be addressed as unique statistical problem it would be an enormous undertaking. However, this is not necessary as it is not the detail of the procedures which is important but rather the type of information produced and the use to which it is put. Therefore, the same principles can be established for analysing results from protein electrophoresis, DNA profiling or gene probing systems or from any similar technique which may be developed in the future. In essence, by considering the type of information produced, all the different procedures can be grouped into two classes which could be termed 'defined' and 'open' systems.

In the defined systems there is a fixed and known number of possible bands and this is associated with an expected genetic interpretation. For example, with electrophoresis such systems are represented by the examination of isozymes in individual plants of perennial ryegrass (*Lolium perenne* L.) by Hayward *et al.* (1976) or of glutenin seed proteins in wheat (*Triticum aestivum* L.) by Payne & Lawrence (1983). On the molecular biology side this situation would probably only apply to the probing for a particular gene construct which had been introduced.

In these systems small changes in band position or intensity do not effect the results. The patterns are interpreted as a whole according to an expectation rather than recorded as exact band positions or intensities. Therefore, in these cases, assuming complete 'repeatability', only biotic variation need be accounted for statistically.

In the open systems there is no predictability or interpretation involved in the results. Unknown numbers of bands can appear and are scored as present at a specific position, given as a 'Rf' value, molecular weight or base pair number and have, on occasion, also been scored for intensity. On the molecular biology side, band matching between samples or gels is also often on a subjective basis and this can lead to problems in matching identities not only in plant science but also in forensics as outlined by Balding & Donnelly (1984).

It is these open systems which show the greatest need for application of the biometricians expertise. Standardised statistical principles need to be established and experimental practices modified to permit the measurement of confidence limits. This will determine the magnitude of difference in band location or intensity which can be accepted as an actual and repeatable genetic difference between two unknown samples. However, even then the use of these types of systems may be more appropriate in the identification rather than the registration role.

Biometricians are also now going to be faced with an additional problem due to the emergence of automatic machine reading of gels by image analysis. This involves all aspects of results handling including the interpretation and analysis of results and the production of 'libraries' of variety-gel identities. Software such as Gelcompar (Anon. 1993) already exists to permit the automatic matching of bands, automatic compressing or stretching of gel lanes and the automatic application of various cluster and principle components analyses to assess the overall genetic difference between samples. Without true genetic interpretation of bands and careful statistical analysis some very extravagant claims based on assumed genetic distance may be made.

These computer systems also facilitate the direct comparison of patterns from different gels, by 'merging' samples common to each gel and then manipulating the traces of unknown samples based on a number of rules, to form computer libraries. Independent statistical guidance is required on the design of testing procedures and upon which analyses should be applied to produce "safe" decisions on variety registration. That is, the number of samples to be taken per variety, the number of replicate extractions and separations used or the number of common samples needed across different gels which are to be merged. These areas all need to be examined before the eventual variety identities, computer libraries and distance analyses can be accepted as valid and sufficiently accurate to permit interpretation by the taxonomist or geneticist.

Discussion

There has undoubtedly been a tendency over the years, certainly in some crops where breeding is very active or the gene-pool small, for testing authorities to take smaller and smaller differences into account for Distinctness. It is of course desirable that the breeder of an improved variety should be able to achieve registration for his innovation. However the responsibility also rests with the testing authority to provide sufficient protection for existing varieties. In this context it is significant that the discussions as to whether or not a candidate variety is "clearly distinguishable" from an already protected variety usually involve a dialogue between the testing authority and the breeder of the new variety. The breeder of the existing variety whose interests must be protected is usually totally unaware of any possible infringement of his right. Testing authorities must therefore take a conservative line in the examination of claimed innovation and arbitrate between on the one hand the need to protect existing intellectual property and on the other the need to reward genuine new innovation. This is a difficult balance which could be upset by an over-eager embracing of all DNA fingerprinting techniques. This is especially so if in future the Essential Derivation concept were to be used to allow a more liberal interpretation of Distinctness.

Plant Breeders' Rights is only of value to the breeder, who, after all, largely pays for the system, if it provides a realistic protection of innovation for a long enough period during which to obtain reasonable recompense, including profit, for R&D expenditure in the breeding of the variety. The concept of "minimum distance", with all its attendant problems, therefore remains an issue for continuing debate and Distinctness cannot be reduced simply to differences in a few nucleotide base-pairs.

Care must be taken that the proliferation of techniques and methods which can identify variation within existing varieties and reveal quite small differences at either the genome, genotype or phenotype level does not lead to a situation where the term "clearly distinguishable" becomes devalued.

It is also important that we do not allow the issues to become clouded by a failure to consider the essential differences between the expressed and non-expressed parts of the genome. What is important is that any difference identified on a gel should have a clearly understood genetic basis and should lead to true phenotypic variation in the plant.

Since the signing of the first UPOV Convention in 1961 the use of phenotypic Distinctness, has quite clearly allowed both polygenic and single gene characteristics to confer Distinctness upon varieties depending upon the perceived 'importance' of the characteristics. History cannot be re-written and therefore this anomaly cannot be removed. Perhaps the only way in which a balance of protecting innovation and yet allowing the opportunity for further advancement can be achieved in the future is for some consideration of minimum (genetic) distance to begin to be included in the determination of whether varieties are "clearly distinguishable" and hence Distinct. This may be subjective and vary according to crop as at present or may have to be given a more formal and perhaps statistical identity in future. Minimum distance could be considered as a minimum difference in or change required in the total genotype of one protected variety before protection can be granted to another similar variety. It probably cannot, however, be considered simply only in terms of quantitative genetics because of the differing economic importance of certain resulting phenotypic differences, for example, the various disease resistance genes.

Genetic distance, using perhaps different scales of differences is probably going to be very useful in consideration of Essential Derivation. However, it must be recognised that this concept, while introduced and described, is not directly encompassed within the UPOV variety protection system. There will probably therefore be a greater need in future for breeders to become more aware of genetic distance for negotiation, and perhaps even litigation, concerning the more complex intellectual property considerations across varieties and genes.

Molecular biology can be of tremendous assistance in the evaluation of genetic distance or minimum distance between varieties whether for Distinctness or Essential Derivation purposes. This potential must not be wasted by short-sighted adoption of methods which simply examine band pattern differences on gels without evaluating true genetic distance. To this end much more development work is needed in the mapping of the markers that are used in various crops to ensure a good distribution across the whole genome. This area must be given much more attention before we step into widespread application of DNA profiling. Simple barcode identification does not meet the requirements of variety registration in either the Distinctness or Essential Derivation situation.

In the application of biotechnological techniques to variety registration there is, as in many areas of science, the requirement for a multi-disciplinary approach. While molecular biologists have made much progress in developing and carrying out the procedures as have the taxonomists and geneticists in interpreting the results, much less emphasis has been given to statistical aspects. It is perhaps timely for biometricians to begin to consider the factors discussed in this paper and clearly establish statistical rules and acceptable operating procedures for assessing Distinctness and measuring genetic distance between varieties.

In conclusion, for *variety identification* the availability of DNA profiling and gene probing techniques will allow significant scientific advances. When the problem is simply determining whether a given seed sample is of the variety stated or even to establish the identity of an unknown sample by comparison with varieties from an established reference collection then the techniques will provide cost effective and rapid methods. However, this is quite different from the situation in *variety registration* where the identity and description of a reputedly novel selection must be established *de novo* and its Distinctness, Uniformity and Stability proven.

The link to genetic interpretation of observed differences in DNA structure and organisation and an understanding of the functional role or phenotypic expression of these differences are two principles which must be adhered to firmly in any application to variety registration and Distinctness.

The use of certain molecular biology methods may therefore be inappropriate but those which satisfy these principles may, alongside more conventional morphological observations, have a significant role to play across the Variety Distinctness, Essential Derivation and Gene Patenting areas within future intellectual property protection systems for plant breeding.

References

Anon., 1978. International Convention for the Protection of New Varieties of Plants - UPOV Publication No. 644 (E) Section 2.

Anon., 1991. International Convention for the Protection of New Varieties of Plants - UPOV Publication No. 644 (E) Section 1.

Anon., 1993. Gelcompar - Comparative analysis of electrophoresis patterns. Applied Maths, B-8511 Kortrijk, Belgium, 4 pp.

Balding, D.J. & P. Donnelly, 1984. How convincing is DNA evidence. Nature 368: 285-286.

Bunce, N.A.C., B.G. Forde, M. Kreis & P.R. Shrewry, 1986. DNA restriction fragment length polymorphism at hordein loci: application to identifying and fingerprinting barley cultivars. Seed Science and Technology 14: 419-429.

Demeke, T., L.M. Kawchuk & D.R. Lynch, 1993. Identification of potato cultivars and clonal variants by random amplified polymorphic DNA analysis. American Potato Journal 70: 561-570.

Hayward, M.D. & N.S. McAdam, 1977. Isozyme polymorphism as a measure of distinctness and stability in cultivars of *Lolium perenne*. Zeitschrift für Pflanzenzuchtung 79: 59-68.

Hu, J. & C.F. Quiros, 1991. Identification of broccoli and cauliflower cultivars with RAPD markers. Plant Cell Rep. 10: 505-511.

Joshi, C.P. & H.T. Nguyen, 1993. Application of the random amplified polymorphic DNA technique for the detection of polymorphism among wild and cultivated tetraploid wheats. Genome 36: 602-609.

Payne, P.I. & G.J. Lawrence, 1983. Catalogue of alleles for the complex gene Loci, Glu-A1, Glu-B1, Glu-D1 which code for HMW subunits of glutenin in hexaploid wheat. Cereal Research Communications 11: 29-35.

Smith, J.S.C., O.S. Smith & S.J. Wall, 1991. Associations among widely used French and US maize hybrids as revealed by restriction fragment length polymorphisms. Euphytica 54: 263-273.

Welsh, J., R.J. Honeycutt, M. McClelland & B.W.S. Sobval, 1991. Parentage determination in maize hybrids using the arbitrarily primed polymerase chain reaction (AP-PCR). Theor. Appl. Genet. 82: 473-476.

Williams, C.E. & D.A. St. Clair, 1993. Phenetic relationships and levels of variability detected by restriction fragment length polymorphism and random amplified polymorphic DNA analysis of authenticated and wild accessions of *Lycopersicum esculentum* L. Genome 36: 619-630.

Wright, C.E., T.J. Gilliland & M.S. Camlin, 1983. Electrophoresis: Implications for Plant Breeders' Rights. Proceedings of ISTA Symposium on Biochemical Tests for Cultivar Identification: 41-50.

Mating system and the effect of heterogeneity and heterozygosity on phenotypic stability

J. Léon, Inst. f. Pflanzenbau und Pflanzenzüchtung, Christian-Albrechts-Universität, Olshausenstr. 40, D-24118 Kiel, Germany

Introduction

Mating systems of crops strongly influence plant breeding procedures. For several crops, breeders have learned to alter the naturally occurring mating system. While naturally inbreeding populations show a high degree of heterogeneity and a low degree of heterozygosity, usually selection procedures result in a homogeneous crop stand. The possibility of producing hybrids from inbred lines turns the plant stands into homogeneous and heterozygous. Breeders, of course, are able to modify the level of heterogeneity and heterozygosity. Practically, the breeder often can choose the type of cultivar (pure line cultivar, population, hybrid), which at the same time is a decision with regard to heterozygosity and heterogeneity level, as well.

Yield stability is becoming more and more important in cropping. Numerous studies have shown, that yield stability is influenced by the type of cultivar or, as mentioned above, by the level of heterogeneity and heterozygosity. For a general understanding it is important to know, which one of these factors possesses the higher effect on yield stability, whether they are equally important or whether interactions between them exist. But besides the breeders ability to influence the mating system of crops and to produce hybrids, the natural mating system still has a great impact on the crop. As an example, outbreeding species like maize and rye usually exhibit higher heterosis levels than inbreeding species like wheat and barley. Therefore, the effect of the mating system on the influences of heterogeneity and of heterozygosity level on yield stability is considered in this study.

During the last decades numerous studies on the methodology of measuring yield stability have been published. It is generally accepted that yield stability is related to genotype by environment interaction. Recently published methods concentrate on multivariate analyses. Since I will review published results including older ones, in this study yield stability is represented by the genotype by environment interaction component of the respective group, the deviation mean squares from the well known regression approach or by ecovalence. In order to compare different crops, and even more important, different levels of heterozygosity (*e.g.*, inbred lines and hybrids) coefficients of variation of the respective measures are calculated whenever possible.

Outcrossing species

Effect of heterozygosity

Natural populations of outbreeders are heterogeneous and highly heterozygous. Comparisons of homogeneous plant stands, *e.g.*, inbred lines and single cross hybrids, make it possible to measure the effect of heterozygosity without simultaneously changing the heterogeneity level. For maize, rye and sunflower these comparisons revealed a tremendous effect of heterozygosity on yield stability (Table 1). All hybrids possess lower stability values than the inbred lines. However, for outbreeders the usual type of cultivar is the population, which is highly heterogeneous and heterozygous. Breeders will not decide between inbred lines and hybrids. The alternative types of cultivars are populations and hybrids. Therefore it is interesting whether the found effect of heterozygosity on yield stability is linear to the inbreeding coefficient. Wahle & Geiger (1978) not only tested inbred lines and their hybrids (Table 1) but also populations and did not observe a further increase in yield stability by increasing heterozygosity level. These populations, of course, were heterogeneous and no exact separation between the effects of heterogeneity and heterozygosity is possible. Rowe & Andrew (1964), as well, did not restrict their analysis on inbreds and F1-hybrids. They also included F2, F3 and the backcross generations BC1 and BC2 in their experiment. These generations, of course, are not homogeneous and no answer can be given on the effect of heterozygosity alone, but their data indicate (Table 2), that homogeneous stands of inbred lines may rather be classified as unstable and that the relationship between yield stability and

Table 1. Effect of heterozygosity on yield stability in outbreeding crops

Crop	Inbred lines		Hybrids		Source
	Yield	Stability*	Yield	Stability*	
Maize	3260	24.3**	6510	9.9**	Rowe & Andrew 1964
Rye	20.7	29.3	65.3	5.1	Wahle & Geiger 1978
Maize	32.1	15.4	81.9	4.3	Böhm & Schuster 1985
Sunflower	5.1	29.5	13.3	17.1	Stamm & Schuster 1985
Maize	24.3	22.8	76.7	6.2	Schnell & Becker 1986

* CV of Deviation mean squares or of Ecovalence, ** CV Genotype by environment interaction

Table 2. Yield stability and level of heterozygosity for maize (Rowe & Andrew 1964)

Inbreds (0%)		BC2 (25%)		BC1 (50%)		F1 (100 %)	
Yield	Stability*	Yield	Stability*	Yield	Stability*	Yield	Stability*
3260	24.3	4470	11.4	5130	12.4	6510	9.9

* CV of genotype by environment interaction

inbreeding coefficient is possibly not linear. The hypothesis that inbred lines of outbreeders are unstable rather than that hybrids of outbreeders are extremely stable, is confirmed by the high numerical values of the coefficient of variation of stability measure compared to all other groups (*e.g.*, inbreds and hybrids of self-fertilising crops and of partial allogamous crops).

Effect of heterogeneity

Although outbreeding results in highly heterogeneous and heterozygous populations and plant breeders can alter the type of cultivar for several species to homogeneous hybrids, only a few reports on the effect of heterogeneity on yield stability under a certain level of heterozygosity exist for outcrossing crops. Schnell & Becker (1986) showed that both, blending maize single cross hybrids and blending maize inbred lines resulted in higher yield stability (Table 3). The benefit from heterogeneity was higher at the homozygous level than at the heterozygous level. The high coefficients of variation of deviation mean squares of inbred lines were remarkably reduced by blending even these relatively unstable inbred lines. Only a slight decrease in stability measure was observed at the heterozygous level by increasing heterogeneity. Another way to analyse the effect of heterogeneity in outbreeding species is to compare different hybrid structures (Table 4). These hybrid structures like single crosses, three way crosses and double crosses, differ in their degree of heterozygosity but the differences in heterogeneity are more important. Most reports on yield stability of hybrid types showed that the stronger heterogeneous double crosses possess the highest yield stability followed by the three way crosses, which were medium in numerical values of yield stability. The homogeneous single

Table 3. Effect of heterogeneity on yield stability in maize (Schnell & Becker 1986)

	Pure stands		Mixtures	
	Yield	Stability*	Yield	Stability*
Homozygous lines	24.0	22.8	24.3	11.4
Heterozygous hybrids	76.7	6.2	76.3	5.9

* CV of Deviation mean squares

Table 4. Effect of heterogeneity on yield stability in outbreeding crops

Crop	Single crosses		Three way crosses		Double crosses		Source
	Yield	Stability*	Yield	Stability*	Yield	Stability*	
Maize		4.37**				1.53**	Sprague & Federer 1951
Maize		8.95**				2.08**	Sprague & Federer 1951
Maize		55.0				34.2	Eberhart & Russell 1969
Maize	65.1	8.5	62.1	6.8	60.3	4.2	Weatherspoon 1970
Maize				98.5**		120.0**	Hühn & Zimmer 1983
Maize	85.5	4.3	81.9	4.71			Böhm & Schuster 1985
Maize	76.7	6.2			77.4	5.4	Schnell & Becker 1986
Sunflower	13.3	17.1	13.9	14.9	12.9	11.3	Stamm & Schuster 1985
Rye			62.1	4.1	60.8	3.1	Becker *et al.* 1982
Maize		24.9**		24.3**			Geiger *et al.* 1987

* CV of Deviation mean squares or of Ecovalence; ** Genotype by environment interaction

crosses showed the lowest yield stability (Table 4). This ranking is valid for all reports with balanced sets of material. The investigations of Hühn & Zimmer (1983) and Böhm & Schuster (1985) did not confirm the general trend. These authors did not analyse balanced data sets. In the case of Hühn & Zimmer (1983) data from the official registration trials in Germany were used. During the time of investigation the number and the proportion of three way crosses increased. A possible improvement in disease- or stress-tolerance of the new cultivars may have overlapped an effect of heterogeneity on yield stability. Besides the general trend, that double crosses are more stable than three way crosses and three way crosses are more stable than single crosses, within these groups great differences among entries occur in yield stability.

Summarising the results for outcrossing species, both factors, heterogeneity and heterozygosity, influence yield stability.

Inbreeding species

Effect of heterozygosity

Only a few reports comparing yield stability of homogeneous hybrids (*e.g.*, single cross hybrids) and their parents are available for self-fertilising crops (Table 5). All these reports are based on balanced data sets. Heterotic effects were present in a range from 2.7 to nearly 40 per cent. Comparing the stability of the parental material to the homogeneous hybrids no evidence occurred that heterozygosity benefits yield stability. The hybrids even showed higher values, or in other words the hybrids were not as stable as the inbreds. This is in contrast to the findings of outcrossing and partial allogamous

Table 5. Effect of heterozygosity on yield stability in inbreeding crops

Crop	Environ-ments	Parents		Hybrids		Source
		Yield	Stability*	Yield	Stability*	
Wheat	3	21.50	8.63	30.02	10.99	Bhullar *et al.* 1977
Wheat	4	16.55	8.76	18.92	10.77	Jatasra & Paroda 1981
Wheat (harvest index)	8	50.31	10.51	49.90	15.23	Chaudhary *et al.* 1978
Turnip rape (inbreeder)	6		4.85**		4.96**	Joarder *et al.* 1978
Wheat	4	65.10	14.4	66.90	15.20	Borghi & Perenzin 1990

* CV of Deviation mean squares or of Ecovalence; ** Genotype by environment interaction

species. The hybrids of self-fertilising species are difficult to produce and in most cited investigations the F1-hybrids were tested in just one row, whereas the parents and further genotypes (*e.g.*, F2, BC) were tested in two or more rows. This possibly results in higher error variances of the hybrid and in non homogeneous error terms and influences the estimation of yield stability parameters. However, Borghi & Perenzin (1990) produced the hybrids by gametocide and tested parents and hybrids in plots. Their data, as well as the other presented reports, did not reveal any increase in yield stability due to heterozygosity. Also, Johnson & Whittington (1977) and Quisenberry & Kohel (1971) found that barley- and cotton-hybrids, respectively, and their parents did not differ in respect to yield stability. Maeng (1984) found for winter wheat that in general, hybrids seemed to be more damaged by severe environmental stresses and diseases than their parents, this might be an explanation for an even lower yield stability of hybrids. In a study with non-balanced material, Carver *et al.* (1987) tested in each of four years several winter wheat hybrids and pure line cultivars at six locations. In two of those four years the hybrids and pure lines did not differ in yield stability. In one year the hybrids were superior and in the last year the pure lines were superior. So no clear results can be deducted. Reviewing these results, there is evidence, that in inbreeding species individual buffering or yield stability can be a property of specific genotypes not associated with heterozygosity as has been pointed out by Allard & Bradshaw (1964).

Effect of heterogeneity

The comparison of pure lines to mixtures of pure lines is a well known design to study the effect of heterogeneity on yield stability. Numerous reports showed, that heterogeneity or populational buffering support yield stability (Table 6). Comparing three-line mixtures to two-line mixtures generally the yield stability increased with the number of components or the degree of heterogeneity (Table 6). Walker & Fehr (1978) tested soybean mixtures with up to 14 components and found, that stability was higher

Table 6. Effect of heterogeneity on yield stability in inbreeding crops

Crop	Pure lines		Two-line mixtures		Three-line mixtures		Source
	Yield	Stability*	Yield	Stability*	Yield	Stability*	
Soybean	1.199	9.94	1.286	7.20	1.304	5.21	Schutz & Brim 1971
Soybean	27.8	15.00	28.3	13.2	28.1	11.30	Walker & Fehr 1978
Oat (group a)	217	2.18	231	1.18	228	1.39	Pfahler & Linskens 1979
Oat (group b)	193	4.56	198	2.95	197	2.58	Pfahler & Linskens 1979
Peanut	3.66	4.29			3.70**	4.17**	Schilling *et al.* 1983
Soybean	2161	7.05	2155	5.89			Bowen & Schapaugh 1989

* CV of Deviation mean squares or of Ecovalence; ** Four-line mixture

for mixtures than for multiple pure stands and tended to increase until mixtures had eight or more components or in other words with the degree of heterogeneity. However, stability was more variable within than among the levels of heterogeneity respective the numbers of components. The effect of mixtures does clearly not only depend on the number of components. Also the degree of similarity of components influences the yield stability of these heterogeneous plant stands. Pfahler & Linskens (1979) tested one group of oat multilines, which met uniformity standards (Table 6, group b) compared to another group of diverse multilines (Table 6, group a). Within both data sets the CV values of stability measures decreased from pure lines to mixtures. This decrease was more obvious (relative 57 % to 64%) in the group of diverse multilines. However, the uniform group showed lower values indicating a higher yield stability, which is a result of the selection process in this particular experiment, since the components of the diverse multilines were selected on the basis of genotypic differences in their variances across environments. In most experiments high yielding cultivars were mixed without any previous information on their combining or mixing ability.

Blending peanut lines, which were selected for similarity in phenotypic expressions, Schilling *et al.* (1983) did not observe a benefit from these multilines in respect to yield stability.

From these results it can be concluded, that heterogeneity generally increases yield stability, however, just a formal mixing of two or more similar components will not increase heterogeneity of plant stands and not be very effective in increasing yield stability.

For self-fertilising species it seems that yield stability depends on heterogeneity and on specific genotypes, and that yield stability is not positively correlated to an increasing level of heterozygosity.

Partial allogamous species

Regarding mating system and yield stability, it is necessary to know as much as possible about the tested material. However, in most reports no comment is given about the particular degree of self- or cross-fertilising. Therefore I used the information given by Fehr & Hadley (1980) to group species.

Effect of heterozygosity

Comparisons of results concerning homogeneous lines and homogeneous single cross hybrids for partial allogamous crops reveal that in all published reports the heterozygous F1 exhibits higher yield stability (Table 7). Therefore it can be stated, that hybrids of partial allogamous crops should in general show a higher yield stability than pure line cultivars of the same crop. On the other hand, the lines did not show high coefficients of variation for the stability measure, so that they can not be classified as unstable as it has been hypothesised for inbreds of outcrossing species. Compared to the self-fertilising and the cross-fertilising crops the partial allogamous species show an increase in yield stability with increasing heterozygosity. The inbreds of partial allogamous crops showed numerical values for coefficients of variation of stability measures, which are comparable to the ones of inbreds of self-fertilising crops. There is evidence, that F1 hybrids of partial allogamous crops benefit from increasing heterozygosity in respect to yield stability without exhibiting a very low stability of their inbreds.

Effect of heterogeneity

For the partial allogamous crops sorghum, rapeseed and faba bean, blending homozygous lines or heterozygous single cross hybrids generally resulted in a higher yield stability of the mixtures (Table 8). That means, that for partial allogamous crops the heterogeneity is an important factor to stabilise yield. From comparisons of single cross to three way hybrids this tendency was also observed, however, this decrease in the coefficient of

Table 7. Effect of heterozygosity on yield stability in partial allogamous crops

Crop	Inbred lines		Hybrids		Source
	Yield	Stability*	Yield	Stability*	
Sorghum	48.2	18.7	60.3	15.5	Reich & Atkins 1970
Sorghum	14.4	7.2	27.1	3.9	Jowett 1972
Sorghum	54.8	12.8	67.1	9.5	Patanothai & Atkins 1974
Sorghum	27.2	5.3**	34.5	2.2**	Rao & Rao 1978
Rapeseed	29.0	10.5	33.3	6.9	Léon 1991
Faba bean	38.1	12.0	53.3	9.2	Stelling *et al.* 1994

* CV of Deviation mean squares or Ecovalence; ** Ratio Deviation mean square to error

Table 8. Effect of heterogeneity on yield stability in partial allogamous crops

Crop	Material	Pure stands		Mixtures		Source
		Yield	Stability*	Yield	Stability*	
Sorghum	Inbreds	48.2	18.7	49.0	14.8	Reich & Atkins 1970
	Hybrids	60.3	15.5	61.4	12.3	Reich & Atkins 1970
Rapeseed	Inbreds	29.0	10.1	30.7	7.0	Léon 1991
	Hybrids	33.3	6.9	35.4	4.6	Léon 1991
Faba bean	Inbreds	38.1	12.0	39.3	8.6	Stelling *et al.* 1994
	Hybrids	53.3	9.2	54.8	5.8	Stelling *et al.* 1994

* CV of Deviation mean squares or of Ecovalence

Table 9. Effect of heterogeneity on yield stability for partial allogamous crops-comparison of hybrid structures

Crop	Single crosses		Three way crosses		Source
	Yield	Stability*	Yield	Stability*	
Sorghum	27.1	3.9	28.3	3.2	Jowett 1972
Sorghum	67.1	9.5	66.5	8.4	Patanothai & Atkins 1974

* CV of Deviation mean squares or of Ecovalence

variation of stability measures was not so obvious as in the experiments of blending two components (Table 9).

For partial allogamous species both factors, heterogeneity and heterozygosity, affect yield stability.

Comparison of the effects of heterozygosity and heterogeneity

Heterozygosity and heterogeneity both affect yield stability. In order to distinguish the importance of each factor, experiments with balanced material containing both factors are necessary. However, only a few experiments are available on this topic and none of them deals with self-fertilising crops. Comparing these experiments, the importance of heterozygosity for outbreeding species is obvious (Table 10). However, as it has been stated earlier, this importance of heterozygosity may be overestimated by comparing inbreds of cross-fertilising species to their hybrids. For both groups, outbreeders and partial allogamous, increasing heterogeneity results in increasing yield stability. For the partial allogamous crops, the increase due to heterogeneity is slightly higher than the increase due to heterozygosity. For maize, the only example of the outcrossing species, the effect of heterogeneity on yield stability was lower than the effect of heterozygosity. But still heterogeneity had a tremendous effect at least on the homozygous level.

In the experiments presented in Table 10, not only the effects of heterogeneity and heterozygosity can be compared but also the interaction between them can be shown.

Table 10. Interaction effects between heterogeneity and heterozygosity for yield stability*

Crop	Inbreds		Hybrids		Source
	Pure stands	Mixtures	Pure stands	Mixtures	
Outcrossing crops					
Maize	22.8	-11.4	-16.6	-17.2	Schnell & Becker 1986
Partial allogamous crops					
Sorghum	18.7	-3.9	-3.2	-6.4	Reich & Atkins 1970
Rapeseed	10.5	-4.6	-3.6	-5.9	Léon 1991
Faba bean	12.0	-3.4	-2.8	-6.2	Stelling *et al.* 1994
Self-fertilising crops					
No experiment available					

* CV of Deviation mean squares or of Ecovalence

From the pattern of interaction effects until now, it is not possible to distinguish between the mating systems. For sorghum and faba bean the yield stability of the F1-hybrid mixtures seems to be additively combined from the heterogeneity and heterozygosity effects, while in rapeseed and the outcrossing maize the joint effect of both factors were only slightly larger than that of the most important factor. Schnell & Becker (1986) characterised this type of interaction of both factors in the sense of "diminishing returns".

Ploidy level and yield stability

Heterozygosity is very important for plant breeders due to the intra-locus interaction of alleles. For autopolyploid organisms, the allelic situations and therefore their intra-locus interaction patterns are much more complex than in diploid organisms. It has been stated that the amount of heterosis depends on the frequent occurrence of more than two alleles per locus and their interaction patterns. In several cases, plant breeders have to decide which ploidy level will fit the breeding goals best (*e.g.*, sugar beet, diploid and autotetraploid rye, forage grasses, etc.). However, there is almost no report on the relationship of ploidy level and phenotypic stability. Pfahler *et al.* (1983) compared in a not balanced material diploid adapted, diploid not adapted and autotetraploid populations of forage rye (Table 11). They found the diploid cultivars to be more stable than the autotetraploid ones. However, the autotetraploid cultivars were not developed in the region of testing whereas the adapted diploids were developed there. Especially for outbreeding autopolyploid crops the effect of heterozygosity and yield stability might be very important.

In regions in which stress factors occasionally are present the proportion of polyploid species is higher than in other regions. Polyploidy is one factor which might be

Table 11. Ploidy level and yield stability for forage rye (Pfahler *et al.* 1983)

Diploid adapted population		Diploid wide crosses		Autotetraploid	
Yield	Stability*	Yield	Stability*	Yield	Stability*
193	1.89	180	1.39	209	2.65

* CV of Deviation mean squares

associated to adaptability. Gupta & Misra (1987) compared four genome combinations of triticum and triticale, respectively, and did not found any relationship from genome combination to yield stability. Similar results have been reported by Quisenberry & Kohel (1971), who tested cotton cultivars of one diploid and two allotetraploid species to determine the effect of ploidy on yield stability. The level of ploidy did not affect the phenotypic stability of the tested characters. The diploid species was as well buffered as the two allotetraploids. Thus the intergenomic heterozygosity associated with amphidiploidy revealed no advantage. Cultivated cotton is self-pollinating. As has been stated for self pollinators, the heterozygosity has no benefit regarding yield stability. This can be the same situation for the intergenomic heterozygosity.

Lessons from micro evolution

The comparison of mating systems in its relevance to yield stability revealed, that outbreeding and inbreeding, conducted for a long time, both result in adapted populations. Both categories of populations are heterogeneous, one highly homozygous, the other highly heterozygous. Plant breeding procedures and demands from farmers and from markets often led to homogeneous crops, which for all mating systems are inferior to the heterogeneous ones in respect to yield stability. In terms of micro evolution, a homogeneous plant stand can not react on changing or on divers environmental conditions. So, from the micro evolutionary point of view and for yield stability, heterogeneity is favourable. Regarding the heterozygosity, the cross-pollinators show high yield stability in the heterozygous state, while self-pollinators, being adapted to homozygosity, show high stability in the homogeneous state, but the inbreds of usually outcrossing species were very unstable. Heterozygosity, therefore, is no prerequisite for high yield stability. We have to consider the mating system first. Allard (1990) studied the process of micro evolution in out- and inbreeding species with regard to adaptation to different environmental conditions. Allard and his co-worker used marker techniques to describe the micro evolutionary changes. For inbreeding species, wild and cultivated as well, they observed clustering of polymorphic loci into groups of three or four loci each by the third or fourth generation after synthesis of a population. A series of breakdowns of associations and complex rearrangements then started that led to clusters of six to eight or more loci in the later generations. The increasing in adaptedness in a given environment was correlated with the development of clusters of associated alleles at different loci. Outbreeding species developed far less genetic structure in the form of these clusters. All of the tested composite crosses of barley developed essentially the same multilocus structure when they were grown in ecologically similar environments. However, they developed very different multilocus structures in each major climatic region in which they were grown. The ecogenetical differentiation among outbreeding populations was much smaller than the ecogenetic differentiation among inbreeding populations (Allard 1990). The reports presented in this study revealed, that inbreeding crops do not need to be highly heterozygous to show a high yield stability. These inbred lines and the respective hybrids do no differ largely in yield stability. It seems that hybridisation disturbs a fine tuned interaction between the developed clusters rather than lead to a further improvement in adaptedness or yield stability, respectively. Outbreeders and probably partial allogamous crops, without these strong clustered groups of associated alleles, respond to increasing heterozygosity with increasing yield stability.

What will happen, if plant breeders change the mating system? Today we do not have exact information on this topic. But the *B. campestris* data (Joarder *et al.* 1978, compare Table 5) provide information. *B. campestris* usually is considered to be a highly outcrossing crop. However, it is well known, that some Indian populations have evolved into inbreeding populations. The data of Joarder *et al.* (1978) with these self-fertilising populations are consistent to all other presented data of self-pollinators and contrary to the data of outbreeders. This indicates that in the long term, changing mating systems will change the relationship between heterozygosity and yield stability. However, we do not have any information about the time or the numbers of generations which are required to alter the relationship of heterozygosity to yield stability.

References

Allard, R.W., 1990. Future directions in plant population genetics, evolution, and breeding. 1-19 In: H.D. Brown, M.T. Clegg, A.L. Kahler & B.S. Weir (Eds.) Plant Population, Genetics, Breeding, and Genetic Resources. Sinauer Associates Inc. Publishers Sunderland, Massachusetts.

Allard, R. W. & A.D. Bradshaw, 1964. Implications of genotype-environmental interactions in applied plant breeding. Crop Sci 4: 503-508.

Becker, H.C., H.H. Geiger & K. Morgenstern, 1982. Performance and phenotypic stability of different hybrid types in winter rye. Crop Sci 22: 340-344

Bhullar, G.S., K.S. Gill & A.S. Khehra, 1977. Stability analysis over various filial generations in bread wheat. Theor Appl Genet 51: 41-44.

Böhm, H. & W. Schuster, 1985. Untersuchungen zur Leistungsstabilität des Kornertrages von Mais (*Zea mays* L.). Z. Acker- und Pflanzenbau 154: 222-231.

Borghi, B. & M. Perenzin, 1990. Yield and yield stability of conventional varieties and F1 bread wheat hybrids. J Genet & Breed 44: 307-310.

Bowen, C.R. & W.T. Schapaugh, 1989. Relationships among charcoal rot infection, yield, and stability estimates in soybean blends. Crop Sci 29: 42-46.

Carver, B.F., E.L. Smith & H.O. England, 1987. Regression and cluster analysis of environmental responses of hybrid and pure line winter wheat cultivars. Crop Sci 27: 559-664.

Chaudharry, B.S., R.S. Paroda & V.P. Singh, 1978. Stability and genetic architecture of harvest index in wheat (*Triticum aestivum* L.). Z. Pflanzenzüchtung 81: 312-318.

Eberhart, S.A. & W.A. Russell, 1969. Yield and stability for a 10-line diallel of single-cross and double-cross maize hybrids. Crop Sci 9: 357-361.

Fehr, W.R. & H.H. Hadley, 1980. Hybridization of Crop Plants. American Society of Agronomy and Crop Science Society of America, Publishers Madison, Wisconsin, USA.

Geiger, H.H., A.E. Melchinger & G. Seitz, 1987. Vorhersage der phänotypischen Stabilität von Dreiweghybriden bei Mais. Vortr. Pflanzenzüchtung 12: 145-155.

Gupta, P.K. & A.K. Misra, 1987. Effect of genome combinations on stability of yield and yield components in wheats and triticales. Theor Appl Genet 73: 899-902.

Hühn, M. & E.W. Zimmer, 1983. Einige experimentelle Ergebnisse zur phänotypischen Stabilität von Doppel- und Dreiweghybriden bei Mais. Z Pflanzenzüchtg 91: 246-252.

Jatasra, D.S. & R.S. Paroda, 1981. Genotype-environment interaction in segregation generations of wheat. Indian J Genetics and Plant Breeding 41: 12-17.

Joarder, O.I., S.K. Ghoes & M. Salehuzzaman, 1978. Genotype-environment interaction shown by yield and some of its components of *Brassica campestris* L. Z Pflanzenzüchtung 81: 248-257.

Johnson, G.F. & W.J. Whittington, 1977. Genotype-environment interaction effects in F1 barley hybrids. Euphytica 26: 67-73.

Jowett, D. 1972. Yield stability parameters for sorghum in East Africa. Crop Sci 12: 314-317.

Léon, J,. 1991. Heterosis and mixing effects in winter oilseed rape. Crop Sci 31: 281-284.

Maeng, D.J., 1984. Studies of yield and yield components in hybrid winter wheat (*Triticum aestivum* L.). Dissertation-Abstracts-International, -B-Sciences and Engineering 44: 11.

Patanothai, A. & R.E. Atkins, 1974. Yield stability of single crosses and three-way hybrids of grain sorghum. Crop Sci 14: 287-290.

Pfahler, P.L. & H.F. Linskens, 1979. Yield stability and population diversity in oats (*Avena* sp.). Theor Appl Genet 54: 1-5.

Pfahler, P.L., H.H. Luke & R.D. Barnett, 1983. Stability parameters for forage production and quality in rye (*Secale* sp.). Z Pflanzenzüchtung 90: 42-55.

Quisenberry, J.E. & R.J. Kohel, 1971. Phenotypic stability of cotton. Crop Sci 11: 827-829.

Rao, S.S. & K.V. Rao, 1978. Genotypic stability of sorghum varieties and hybrids. Indian J agric Sci 48: 691-695.

Reich, V.H. & R.E. Atkins, 1970. Yield stability of four population types of grain sorghum, *Sorghum bicolor* (L.) Moench, in different environments. Crop Sci 10: 511-517.

Rowe, P.R. & R.H. Andrew, 1964. Phenotypic stability for a systematic series of corn genotypes. Crop Sci 4: 563-567.

Schilling, T.T., R.W. Mozingo & T.G. Isleib, 1983. A comparison of peatnut multilines and component lines across environments. Crop Sci 23: 101-105.

Schnell, F.W. & H.C. Becker, 1986. Yield and yield stability in a balanced system of widely differing population structures in *Zea mays* L. Plant Breeding 97: 30-38.

Schutz, W.M. & C.A. Brim, 1971. Inter-genotypic competition in soybeans. III. An evaluation of stability in multilines mixtures. Crop Sci 11: 684-689.

Sprague, G.F. & F.T. Federer, 1951. A comparison of variance components in corn yield trials. II. Error, year × variety, location × variety and variety components. Agron J 42: 535-541.

Stamm, U.I. & W. Schuster, 1985. Leistungsstabilität verschiedener Hybridtypen bei der Sonnenblume. Ber

Arb-Tag AG Saatzuchtleiter, Gumpenstein: 219-228.

Stelling, D., E. Ebmeyer & W. Link, 1994. Yield stability in faba bean, *Vicia faba* L. Plant Breeding 112: 30-39.

Wahle, G. & H.H. Geiger, 1978. Vergleich der phänotypischen Stabilität von Inzuchtlinien, Einfachhybriden und Populationen bei Winterroggen. Z Pflanzenzüchtung 80: 211-222.

Walker, A.K. & W.R. Fehr, 1978. Yield stability of soybean mixtures and multiple pure stands. Crop Sci 18: 719-723.

Weatherspoon, J.H.,1970. Comparative yields of single, three-way, and double crosses of maize. Crop Sci 10: 157-159.

Marker-assisted breeding

Piet Stam, Centre for Plant and Reproduction Research, CPRO-DLO, PO Box 16, 6700 AA Wageningen, and Department of Genetics, Wageningen Agricultural University, Dreijenlaan 2, 6703 AH Wageningen, The Netherlands

Key words

linkage analysis, QTL mapping, marker-assisted selection, gene tracing

Abstract

Various aspects of the use of molecular markers in plant breeding are discussed. Attention is given to some general problems arising in linkage analysis with molecular markers. Some limitations and prospects of QTL mapping are briefly sketched. Attention is further paid to the question of how genetic information, obtained from single gene and QTL mapping, can be exploited in practical plant breeding. The prospects of marker-assisted selection, "gene tracing" and cross prediction are briefly reviewed. Some problems that may arise when using DNA markers in variety registration are addressed. Finally, needs for transferring current fundamental knowledge to practical plant breeding are outlined.

Introduction

The potential use of marker genes for genetic analysis of quantitative variation has already been recognized in the twenties and thirties of this century. Sax (1923) and Rasmusson (1932) demonstrated the existence of linkage association between qualitative and quantitative traits in crosses between genetically different stocks. Only when large numbers of molecular markers became available, the idea that these could provide a useful tool in breeding plants and animals, has received renewed attention. The advent of the various techniques to uncover variation at the DNA level, and the detection of the overwhelming amount of polymorphism at the DNA level has initiated a revival of biometrical genetics. The notion that molecular markers may provide a way to identify and map the genes controlling quantitative variation has greatly stimulated the development of new statistical methods. Thus far, quantitative genetics had only been able to describe and analyze genetic variation in terms of statistical concepts, *i.e.*, means, variance components, correlations, etcetera. Although several methods had been suggested to estimate the number of genes involved in quantitative characters, these methods were known to be of little value outside their own theoretical context, and therefore, the individual genes underlying quantitative traits were assumed to exist, but only their joint effects were subject of analysis.

Very soon, the early advocates of using molecular markers envisaged the genome of crop species as being covered with numerous tags, each tag labelling an important gene. Today, the various genome mapping projects in plants are indeed pacing towards this state of affairs at a rate increasing by the year.

With respect to application in plant breeding, the molecular marker technology is promising indeed. Very generally stated, marker-assisted plant breeding takes advantage of (linkage) association between markers and agronomic traits in crop species. The advantage for practical breeding may relate to several phases of breeding: from germ plasm management and cross prediction to indirect selection and variety registration. It is now clear that there are many potential applications of molecular markers; however an overall picture of its cost effectiveness, the ultimate criterion for practical breeding, can hardly be sketched at this time.

Linkage associations

Linkage association is one of the key concepts to marker-assisted breeding. Basic population genetics theory tells that little of such associations are to be expected in large outbreeding populations, even for tightly linked genes. Association, or linkage disequilibrium, is to be found in the offspring of crosses between genetically different stocks. Crosses between true breeding lines of autogamous species, crosses between inbred lines and the offspring of single pair matings of allogamous species will display linkage disequilibrium, and are therefore the types of populations in which associations are to be assessed. However, not only segregating generations, also "simple" mixtures of genetically different stocks will display associations. Scanning such mixtures may reveal markers that correlate with agronomic traits, although the correlation may not necessarily reflect genetic linkage.

Once linkage associations have been assessed in a given cross, the aggregate marker genotype can, in various ways, be used as a predictor of genotypic value, breeding value and/or combining ability. With respect to the assessment step, several questions arise;

some of these relate to the accuracy of (classic) linkage estimation, some focus on the analysis of quantitative variation, and some apply to both aspects.

Population type

A variety of segregating offspring generations can be used for linkage analysis. Most commonly used are the classic types: first generation backcross (BC1), F2, and full-sib families (allogamous species). Where possible, doubled haploids (DH), either obtained from F1 or from F2, are attractive options. Also recombinant inbred lines (RIL), preferably of an advanced generation, are being used frequently. For a give size, not every type of population is equally informative for linkage detection. The question of informativeness in this respect is roughly equivalent to the question of how many informative gametes can be traced back from the zygotes.

A great advantage of DH lines and RILs is their "eternity": they can be multiplied generatively without loosing genetic identity. This is especially important for QTL detection, because it enables one to use replicates and to collect observations in multiple environments.

Types of molecular markers

With respect to the type of markers, the main distinction for the purpose of linkage analysis is between dominant and codominant ones. Dominant markers are generally less informative than codominant ones. For the "eternal" RILs and DH lines, which are completely homozygous, both types are equally informative. Apart from their informativeness, the choice of a molecular marker type will also depend on other factors, for example those that relate to the laboratory techniques involved and the degree of polymorphism they can detect. RFLPs, RAPDs, microsatellites and other tagged DNA sequences are all different in this respect.

Complications in linkage analysis

The large amount of linkage data that has become available through molecular markers has caused a revival of the interest in linkage analysis. Most of the statistical theory in this field dates back to the fifties and sixties (Mather 1951, Allard 1956, Bailey 1961). Molecular marker data, however, may cause complications that are not completely covered by this theory.

First, there is the complication of having more than one type of segregation in a single cross. This may happen, for example, in single pair matings of outbreeders. Some markers may, in this case, segregate in a 1 : 1 ratio (backcross type), whereas others may segregate in a 1 : 2 : 1 or 1 : 3 ratio (F2 type), while still others may segregate in a 1 : 1 : 1 : 1 ratio. Also in a classical F2, a complication may arise when in some individuals a marker is scored as codominant (clear presence or absence of a band), whereas in other individuals the same marker is scored as dominant (in case of doubt for one of a pair of allelic bands). Several computer packages for linkage analysis are now available that properly deal with these complications.

Missing data are a potential source of error in linkage analysis of molecular data. Most molecular marker data are scored by presence or absence of a band on a gel or autoradiogram. In case of doubt, as on a poor gel with a high level of background signal, or very faint bands, the genotype is usually scored as "unknown", *i.e.*, as a missing data point. However, the probability of being scored as "unknown" is generally different for the genotypic classes, and therefore missing observations very often are not random. This in turn may lead to erroneous estimates of recombination frequencies.

Mapping quantitative trait loci

In detection and genetic mapping of quantitative trait loci (QTL) there has been a continuous development and improvement of the statistical methods during the last decade. R.C. Jansen's contribution to these proceedings gives an overview of the various approaches that have been taken. Several other contributions are dealing with specific problems, and as such, they witness the current progress in this area.

It is now generally recognized that the most widely used method of "interval mapping", proposed by Lander & Botstein (1989) is unable to deal with two important problems: (a) the detection of linked QTLs, and (b) the detection of QTL by environment interaction. Although the method of interval mapping may in some cases indicate multiple linked QTLs and/or genotype by environment interaction, the picture is in most cases not clear and most likely leaves much behind the screen. The general framework developed by R.C. Jansen is very promising in this respect (*cf.* Jansen 1993, Jansen 1994, Jansen & Stam 1994).

A question addressed frequently in the discussion on QTL mapping is: what is the optimal density of the genetic marker map for QTL mapping, if any? Increasing the number of markers beyond a given average density makes little sense if population size is not increased at the same time. This is because QTL mapping, as well as ordinary gene mapping, relies on the frequency of detectable recombination events, which beyond a given marker density, can only be increased by studying larger populations.

Connected to this problem of accuracy of QTL detection is the accuracy of the marker map that is being used. In most QTL mapping procedures the marker positions are treated as fixed, *i.e.*, as known with absolute certainty. This is, of course, never the case. Marker positions, as well as QTL positions, always represent the best statistical estimate; they are prone to simultaneous and correlated errors, when obtained from the same data.

In a number of studies (Schön 1993, Stuber *et al.* 1992) it has been observed that a given putative QTL, detected in one cross is not "expressed" in another cross. This may be due to epistatic interaction (the difference between the QTL genotypes depends on the genetic background) or simply to the fact that in the "silent" cross no detectable difference between the QTL genotypes exists. The latter may result from multi-allelism at the QTL. When "QTL activity" can not generally be extrapolated from one cross to the next, this may seriously hamper some applications in breeding programs; it would mean that for the purpose of indirect selection via markers, associations must be assessed for every cross anew. The contributions of Charcosset *et al.* and Rebaï *et al.* in these proceedings describe an approach to QTL detection and mapping in which information from several crosses is used simultaneously. It can, in principle, detect multiallelism at QTLs, and may give an indication as to the most promising combination of parents, as long as the set of crosses is connected to a certain degree.

QTL detection and mapping have often been considered as a way to "dissect" complex traits into Mendelian factors (Paterson *et al.* 1991). Especially with respect to agronomically important traits, such as yield, one should always be aware that a putative QTL for the target trait may in fact be a QTL for a correlated trait or a component of the complex target trait, *e.g.*, lateness, which influences the target trait (*e.g.*, yield) in one way or another. At the same time, a QTL analysis for several correlated traits may reveal part of the mechanism behind such correlations. A documented example is given in the contribution by Lindhout *et al.* in these proceedings. Here it is demonstrated that in the tomato cross under consideration, most of the QTLs affecting earliness, also affect fruit size. Of the three earliness-QTLs, one was associated with ripening time, one with fruit set and one with flowering time.

Marker-assisted selection

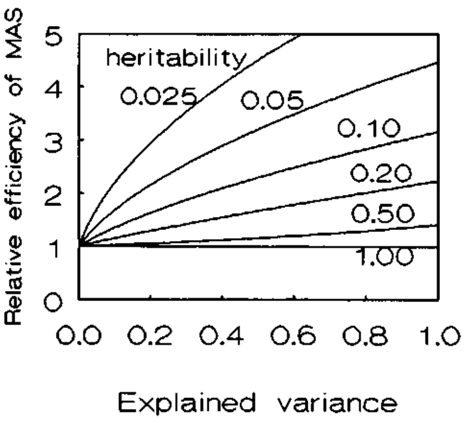
Marker based index selection

Lande & Thompson (1990) have studied the efficiency of marker based index selection, relative to phenotypic selection. They proposed to construct an index in which phenotype and marker genotype are given optimal weights. The relative efficiency, in terms of expected response to selection is given by

$$RE = \sqrt{\frac{p}{h^2} + \frac{(1-p)^2}{1-h^2p}}$$

where h^2 is the heritability of the trait and p is the proportion of the additive genetic variance explained by the markers. A graphical representation of the above equation is given in Figure 1.

Figure 1. The relative efficiency of marker based index selection. From Lande and Thompson, 1990



The main conclusion to be drawn from this result is: the lower the heritability, the more it pays to use marker information. Lande & Thompson were considering an outbreeding model population, with a given amount of initial linkage disequilibrium. In such a population, but also in (partially) inbreeding populations, linkage disequilibrium will decay, and, as a consequence, the markers will explain less and less of the variation as generations proceed. For this reason Lande & Thompson suggested that re-assessment of the marker-trait associations might be required for this form of marker-assisted selection to be competitive with phenotypic selection. This requirement of re-assessment raises the question about cost-effectiveness of the procedure. An important factor in this respect is the size of the population in which associations are assessed. The larger the population, the more accurately trait-marker correlations can be estimated. The contribution by Gallais & Charcosset in these proceedings deals this problem: it is shown that there is a pay-off between heritability and population size in this respect.

In interpreting the theoretical results about the prospects of marker-assisted selection it should be kept in mind that heritability, which is a crucial factor, can in many cases be augmented, if really needed. Using replicates or corrections for environmental heterogeneity are, among others, ways to achieve this.

Marker based index selection does not require that either markers or putative QTLs be mapped. An array of regression coefficients, preferably to be obtained with some regressor selection procedure, is used for computation of the index, the selection criterion. In doing so, much of the genetic information present in the assessment data is neglected. Marker based index selection is another form of indirect selection, exploiting genetic correlations, without taking advantage of the known underlying genetic architecture. In this sense, index selection is a statistical procedure, that is blind to the (putative) Mendelian factors underlying it. Fully exploiting the currently available methods for QTL detection, *i.e.*, locating the genomic regions containing the genes involved in the target trait, seems more promising. Selection procedures that aim directly at the genes underlying quantitative traits can change marker-assisted selection from "blindfolded" statistical selection to real application of Mendelian genetics. In the latter, tagged genome regions are traced in a breeding and selection program.

Gene tracing and genotype construction

Marker-assisted "gene tracing" relies on the possibility to trace labelled chromosome segments in successive generations of a breeding program. This not only involves the selection of favourable alleles or combinations of alleles, it also applies to making a deliberate choice of parents for further crossing, so as to generate new promising combinations of marked chromosome segments.

The most obvious application of "gene tracing" is marker-assisted introgression. Here markers are primarily used to identify that part of the donor genome that does not contain the gene(s) to be introgressed. It has been demonstrated repeatedly (*e.g.*, Chyi *et al.* 1994) that the dilution of the donor genome in a repeated backcross program can be accelerated twofold, compared to random backcrossing. If the map location of the markers is also known, this procedure can be optimized further by a deliberate choice of markers, such that they cover the whole genome, especially the region containing the gene(s) to be introgressed.

Accumulation of favourable alleles is another form of gene tracing. Figure 2 shows a hypothetical, though not unrealistic, result of a QTL analysis, which could serve as a starting point to a gene stacking procedure. Here one can take full advantage of the estimated map positions of the putative QTLs in order to generate the most promising

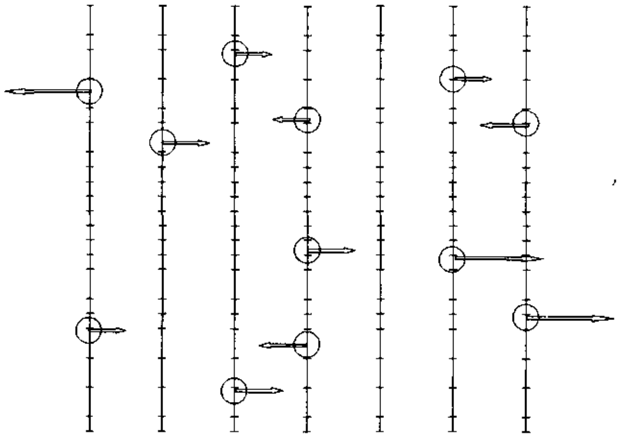
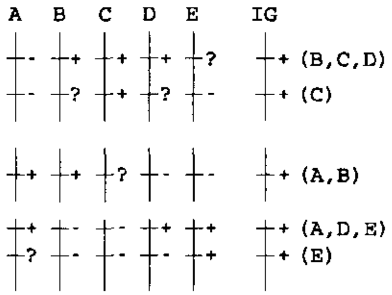


Figure 2. Hypothetical genome with 12 putative QTLs. Arrows indicate direction and size of the effects. Notice that favourable alleles at QTLs were dispersed in the parents

genotype (the ideal genotype would have all the arrows pointing in the same direction). It is easy to verify by calculation that this supposedly ideal genotype will never occur in an F2 or set of RILs of realistic size. However, since the approximate map positions of the QTLs are known, it is possible to identify those individuals which, upon further crossing, are most likely to produce this ideal genotype. Also stepwise procedures (three-way or four-way crosses) can be considered. This approach is a way to breed the most transgressive segregant, applying Mendelian genetics rather than phenotypic selection.

A similar approach to gene stacking can be taken in those situations, where QTL information is available from several crosses. In that case favourable alleles from a number of different origins are to be stacked into one (hopefully) ideal genotype. Figure 3 illustrates a possible, simplified configuration of favourable alleles distributed over five

Figure 3. Hypothetical chromosomal distribution of favourable (+) and unfavourable (−) alleles over five parents (A,..,E). Question marks indicate lack of information. IG: ideal genotype



parents. Also in this case it is not extremely difficult to determine the optimal crossing scheme to arrive at the ideal genotype.

The ideas sketched above are attractive for several reasons. First, they are generally applicable; they apply to any (diploid) crop species, to any trait, and any combination of traits. Second, application in practical plant breeding requires no special or sophisticated techniques. It does require powerful statistical procedures and software for QTL detection.

An area in plant breeding for which gene tracing seems be to very promising is breeding for accumulated partial resistance to pests and diseases. It is generally believed that accumulation of partial resistance factors contributes to durability of resistance, because accumulated resistance genes may constitute a barrier that is hard to knock down by pathogens. When genes for partial resistance from distinct sources are accumulated, their individual effects are hard to be measured from the phenotype. However, when being tagged with markers, their accumulation can proceed beyond the level where an additional gene has no measurable phenotypic effect.

Prediction of heterosis and transgression

The choice of parents to be used in conventional breeding is often considered as the most crucial step in a breeding program. If breeders would have a solid base to predict, for a given set of potential parents, which combinations are most promising, they would have a most valuable tool. A combination of the parental values *per se* and their genetic distance might provide such a means.

With respect to heterosis and its prediction based on genetic distance most of the presently available evidence comes from maize. In maize, the general relation between genetic distance, based on molecular markers, and heterosis is hard to establish because of the existence of heterotic pools. Across such pools heterosis is poor, despite the larger genetic distance. Disruption of coadapted gene pools when crossing between pools, could be the cause of this.

The general existence of a similar phenomenon in autogamous crops is less likely. Predicting transgressive segregation from genetic distance for quantitative traits in inbreeding crop species should be possible if the level of marker diversity parallels the gene diversity at trait loci among a group of potential parents. This parallel in turn depends on the degree of association between marker alleles and QTL alleles in such a set of parents; it might vary from set to set. Conclusive results with respect to these questions have not yet been obtained. The prospects of predicting the performance of parental combinations from genetic distance, inferred from markers, therefore, deserve further attention by means of experimentation and verification.

Other information from markers

The genome mapping projects that are currently being carried out in various crop and other plant species generate a large amount of genetic information that sooner or later may prove to be useful to plant breeding. The detailed genetic maps of crop species that are now available not only represent the positions of DNA markers, they also contain many known genes. This can be helpful to breeding in several ways. First, the phenomenon of conservation of gene orders across species can be exploited. Knowing where a gene maps, *e.g.*, in tomato, makes it easier to spot the corresponding gene in potato when gene and marker orders on the chromosomes are blockwise conserved across related species. Secondly, in cases where a putative QTL has been located in a certain region of the genome, and several genes of known morphological and/or physiological effect also map to that region, this is a good reason to look for possible candidates among these genes that might correspond to the putative QTL. Thirdly, it has been found in several crops that functional clusters of genes may occur in clusters on the genetic map. Examples of this are disease resistance genes in maize, tomato and lettuce (Michelmore 1993, Hu & Hulbert 1994). The intriguing aspect is that such a cluster on the genetic map may contain a variety of genes conferring resistance to biologically diverse pathogens, such as fungi and viruses. Hu & Hulbert (1994) reported of recombination within a complex locus for rust resistance in maize, leading to a race-nonspecific allele.

Although these findings do not immediately contribute to more efficient plant breeding, they are contributing to our understanding of the genetics of agronomic traits, and as such they constitute welcome additional information.

Variety registration and identification

At the end of the plant breeding process, molecular markers may provide a useful instrument with respect to variety registration and identification of breeding lines, identification of contaminated seed lots, etcetera. It is beyond any doubt that DNA fingerprinting is a powerful tool for identification purposes. However, with respect to the protection of breeders' rights and registration of new varieties, a debate seems inevitable. Especially with respect to the concept of "essentially derived varieties" and the use of DNA fingerprinting therein, the picture has not yet been fully drawn. It has been suggested that a certain degree of "concordance", based on DNA markers, between two (claimed) varieties can be used as a basis for a lawful decision about derived versus non-derived. The question in these matters is not so much whether or not one is able to tell the difference between two varieties on the basis of their DNA profile. More fundamental is the issue whether or not neutral DNA markers should be used for this purpose. It should be kept in mind that the legislation on breeders' rights serves a public interest, *i.e.*, the stimulation and encouragement of breeding activities so as to improve crops. Too much emphasis on distinctness of varieties, based on DNA profiles, may endanger this public interest.

Since a number of fundamental biometrical and population genetical problems are to be faced in international legislation and the shaping of international conventions (UPOV), biometricians and population geneticists should play their part in it.

Perspectives, needs for the future

Molecular markers have convincingly entered plant breeding research. During the last five years significant steps forward have been made, not only with respect to marker technology in the laboratory (such as the introduction of PCR), but also in the area of molecular data analysis and gene mapping. Primarily, markers are a useful and powerful tool for genetic analysis. Genetic mapping of single genes (monogenic traits) and resolving quantitative traits into Mendelian factors has not only become within reach, it has been demonstrated to be a reality. Understanding the genetics of agronomic traits is always helpful to the plant breeder. For that reason alone, molecular markers have conquered a firm position in plant breeding research.

With respect to their use in practice, the potential applications cover all phases of plant breeding: from scanning and management of germ plasm to variety registration and identification. However, many of the applications still require a gap between theory and practice to be bridged. Implementation of marker technology on an industrial scale will demand a serious re-designing of breeding programmes. First, the large amount of molecular data that is being generated for the major crops, requires the availability of (adapted) computerized data storage and management systems. Existing data bases will have to be re-shaped so that molecular marker information at any level (probes, probe/enzyme combinations, primer sequences, primer combinations, fragment lengths, genetic marker maps, etc.) can be incorporated and linked to the usual agronomic and pedigree information. Secondly, reliable statistical and computational tools for the processing of marker information are needed. These would comprise genetic mapping, QTL mapping and prediction procedures. Preferably these should be implemented in software packages, either as stand-alone programs or as modules in widely used general statistical packages. Sophisticated computation in this area has been applied so far mainly at laboratory scale; apart from computer packages for "ordinary" genetic marker mapping, virtually no tailored software is publicly available at present. *Ad hoc* procedures, based on a good sense of genetics and biometry can carry the breeder a long way on the road to marker-assisted breeding. However, if the optimistic view, based on the small scale successes, such as reported in these proceedings, is to be validated, considerable efforts are required in quantitative genetics, statistics and the development of software.

Concerning cost-effectiveness of marker-assisted breeding, it is hard to set general guidelines. The diversity of crops with respect to mode of propagation, generation time, crossability with relatives, etc., most likely requires that crop-tailored model calculations be made to obtain a reliable picture of cost-effectiveness. A complication with such model calculations is that genetic gain per unit cost is not the only thing that counts. Squeezing a breeding program into a shorter period of time, *e.g.*, by introducing an extra fast cycle generation per year in which marker-based selection is performed, may pay very well, even if the general agronomic worth of the product (a new variety) is slightly less than what could have been achieved with more effort.

In conclusion, it is not exaggerative to state that the perspectives and possibilities of marker-assisted breeding are numerous, but that a variety of challenges for population and biometrical geneticists still lie ahead, especially for transforming the current methodology into tools for applied breeding.

References

Allard, R.W., 1956. Formulas and tables to facilitate the calculation of recombination values in heredity. Hilgardia 24: 235-278.

Bailey, N.T.J., 1961. Introduction to the mathematical theory of linkage. Clarendon press, Oxford.

Chyi, Y-S, J. Taylor, K. Kellesvig & L. Sernyk, 1994. Results of early versus late selection in backcross breeding using RFLP markers. Abstracts Plant Genome II Conference, San Diego, USA, January 1994, p. 23.

Hu, G. & S. Hulbert, 1994. Recombination at a complex rust resistance locus can generate alleles with race-nonspecific resistance and disease lesion mimic phenotype. Abstracts Plant Genome II Conference, San Diego, USA, January 1994, p. 36.

Jansen, R.C., 1993. Interval mapping of quantitative trait loci. Genetics 135: 205-211.

Jansen, R.C., 1994. Controlling the type I and type II errors in mapping quantitative trait loci. Genetics: in press.

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

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

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

Mather, K., 1951. The measurement of linkage in heredity (2nd edition). Methuen, London.

Michelmore, R.W., 1993. Molecular markers in the manipulation of disease resistance. Abstracts XVIIth International Congress of Genetics, Birmingham, U.K., August 1993.

Paterson, A.H., S. Damon, J.D. Hewitt, D. Zamir, H.D. Rabinowitch, S.E. Lincoln, E.S. Lander, and S.D. Tanksley, 1991. Mendelian factors underlying quantitative traits in tomato: comparison across species, generations and environments. Genetics 127: 181-197.

Rasmusson, J.M., 1933. A contribution to the theory of quantitative character inheritance. Hereditas 18: 245-261.

Sax, K., 1923. Association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris.* Genetics 8: 552-560.

Schön, C-C., 1993. RFLP mapping in maize (*Zea mays* L.): quantitative trait loci affecting testcross performance of elite european flint lines. PhD thesis University of Hohenheim.

Stuber, C.W., S.E. Lincoln, D.W. Wolff, T. Helentjaris & E.S. Lander, 1992. Identification of genetic factors contributing to heterosis in a hybrid from two elite maize inbred lines using molecular markers. Genetics 132: 823-839.

The design of variety trials

M. Talbot, Scottish Agricultural Statistics Service, University of Edinburgh, Edinburgh EH9 3JZ, U.K.

Key words
experimental design, alpha-lattice designs, neighbour designs, g × e variation

Abstract
Variety trials are one of society's primary mechanisms for ensuring that good varieties are identified and put into agricultural use as efficiently as possible. The efficiency with which this transfer takes place depends critically on the effectiveness of the trials system. Here are reviewed some of the recent developments in within-trial design including row-and-column designs, latinized alpha designs, and also some work in progress on designs to minimize interplot interference. The role of neighbour methods vis-a-vis block design-and-analysis methods are assessed.

The design of a trial series is considered and the information provided by components of variance studies of genotype × environment data is explored.

Introduction

The basic principles of experimental design are founded on replication, randomization and local control, *i.e.*, minimizing the effects of experimental variation. In variety testing the best comparison of a set of varieties comes from sowing varieties into a compact block of small plots where soil variation is likely to be minimal. For small numbers of varieties (say up to 15) randomized complete block designs are usually satisfactory and for larger numbers of varieties a range of incomplete block designs have been available in the experimental design literature. These are collectively referred to as lattice designs *i.e.*, simple, triple; square, rectangular; cubic. (Cochran & Cox 1957). Lattice designs all have in common the use of compact smallish blocks any one of which contains only a proportion of the total entries. Thus variety and block effects are confounded and variety means must be adjusted, but the local control so achieved can reduce the error and enhance the precision of variety comparisons.

Alpha designs

A serious drawback of classical lattice designs is that there are constraints on the numbers of varieties and replicates that can be accommodated. In the case of square lattice designs with v varieties it is a condition that v must be a perfect square with $\sqrt{v}$ plots per block. To overcome some of these constraints Patterson & Williams (1976) devised a new class of incomplete block designs called alpha designs. Alpha designs are in some respects a generalization of Dr Frank Yates' original lattice designs. The main advantage of alpha designs is flexibility; they are available whenever the number of varieties is a multiple of the block size, and they can be easily adapted even when it is not.

These early alpha designs were aimed primarily at controlling variation down columns of plots in the field. This is often adequate when plots are long and narrow. Patterson & Hunter (1983) have demonstrated the value of alpha designs in such circumstances. However, when the plots are squarer in shape then designs which allow for both row and column variation can be more effective. Recent developments in design construction (Nguyen & Williams 1993) have shown how alpha designs can be used to produce efficient row-column designs, *i.e.*, where the plots of each replicate are laid out in a rectangular array made up of columns (blocks) and rows (plots per block). Row-column designs are of particular benefit in situations where it is natural to organize plots within a replicate in a rectangular arrangement. The properties of alpha row-column designs are determined by those of the two component block designs, one with blocks given by the rows and the other with the blocks given by the columns, as well as by the way in which the components are put together.

Often in experimental layouts, replicates are placed next to each other so that the columns (blocks) of each replicate form long columns running down the replicates. Then there is a need to ensure that a variety occurs only once in each column. Latinized designs have this property and can be used in conjunction with alpha designs and row-column designs. An example of a situation where latinized designs are useful has been given by Williams (1986).

Designs to minimize interplot interference

Interference between neighbouring varieties has long been recognised as a potential source of bias in estimates of relative variety performance. Factors associated with interference included plant height and disease resistance. Taller varieties dominate smaller neighbours enhancing their own yield artificially and depressing the yields of their neighbours. Disease susceptible varieties export inoculum to their neighbours so that their resistance is over-estimated and that of partially resistant varieties greatly under-estimated.

Prior information is often available on varieties which, when trials are sown, allows them to be grouped so that interference is a minimum between pairs of varieties within a group. In practice, where interference is known to be a problem, varieties from different interference groups are allocated to separate experiments, or to separate blocks in a spilt plot design. However, if the trial site is heterogeneous, variety comparisons between varieties in different groups will be made with poor precision. An alternative is to use a connected block design which reduces the number of times that varieties from different groups occur in the same block. Interference can also be reduced by using neighbour designs where randomizations within blocks is restricted so that adjacent pairs of varieties come from the same or similar interference groups.

As an example of the neighbour method, consider a complete block of 8 varieties divided into four ordered interference groups of size 2, *e.g.*, group 1 may contain the tallest varieties and group 4 the shortest varieties. First, four small groups are defined as follows:

1 1   2 2   4 4   3 3

where varieties have been replaced by their group numbers. Two large groups are then defined:

11   2 2   1 1   2 2

Now randomize large groups within blocks, small groups within large groups, and plots within small groups. This provides eight arrangements of the interference groups:

1	3	4	2	2	4	3	1
2	1	3	4	4	3	1	2
4	2	1	3	3	1	2	4
3	4	2	1	1	2	4	3

Interference is reduced as varieties in groups 1 and 4 do not occur as neighbours.

Furthermore, with this randomization, each variety has an equal probability of appearing of in each plot.

This procedure may be implemented for alpha designs as follows. Consider the case of 20 varieties which are allocated to four equal-sized interference groups. Codes 1-5 correspond to varieties of group 1, codes 6-10 to varieties of group 3, codes 11-15 to group 4 and codes 16-20 to group 2. The layout for the first replicate is:

Block	Variety				Group			
1	1	6	11	16	1	3	4	2
2	2	7	12	17	1	3	4	2
3	3	8	13	18	1	3	4	2
4	4	9	14	19	1	3	4	2
5	5	10	15	20	1	3	4	2

and further replicates have the same pattern of groups. Randomization proceeds as follows:

i) randomize blocks within replicates;

ii) randomize varieties within blocks using groups;

iii) randomize codes to varieties within groups.

Randomization can be further restricted to avoid varieties from groups 1 and 4 appearing as neighbours across blocks. Reduction to three groups can be done by combining groups 2 and 3.

An alpha design randomized as a neighbour design may be analysed with the block structure replicate/block/large group/small group/plot. This provides a valid analysis when the small groups are of equal size. However, the case of unequal blocks needs further study. Neighbour-restricted alpha lattice designs achieve the aim of minimising interference while still providing reasonably precise estimates for all variety differences, although further study is required of the underpinning randomisation theory.

Neighbour methods of analysis

In blocking we group plots so that the average yield differences between blocks, as a result of fertility, etc., can be removed in the subsequent incomplete block analysis. However, soil patterns generally change in a gradual way across a field. In neighbour analysis the plot variation is assumed to be composed of a trend, which changes smoothly from plot-to-plot, plus an independent measurement error.

Many approaches have been proposed for modelling trend by neighbour methods in plot trials, *e.g.*, Papadakis (1937), Bartlett (1978), Wilkinson *et al.* (1983), Patterson & Hunter (1983), Green *et al.* (1985), Williams (1986), Besag & Kempton (1986), Gleeson & Cullis (1987), Lill *et al.* (1988), Martin (1990), Cullis & Gleeson (1991), Baird & Mead (1991). Most methods are based on a 'trend + measurement error' model. The most general, and best developed, approach is that of Cullis & Gleeson (1991) which has been implemented in the computer program TwoD (Gilmour 1992). TwoD provides for the fitting of a range of neighbour models based on the autoregressive integrated moving average (ARIMA) algorithm. As the name indicates the program will fit neighbour models in two directions, along a row, *i.e.*, a bank of plots placed side-by-side, and along a column, *i.e.*, with the plots running end-on down the field.

If plots are long and/or there are only two or three banks of plots then a one-dimensional analysis usually suffices. If the plots are short and there are four or more banks then a two-dimensional analysis is recommended. The sequence in which the model dimensions are fitted will affect the final result but it is general practice to fit first the direction with longest dimension. In the analysis treatment means are removed before fitting the autoregressive terms.

The choice of appropriate neighbour model requires judgement. A sequential strategy is adopted starting with a simple model and fitting models of an increasingly high order until a satisfactory fit is achieved. An important consideration in the selection of an ARIMA model is the level of differencing required to obtain a stationary sequence, *i.e.*, a sequence of adjusted plot values containing no trend. In first differencing the value for each plot is adjusted by subtracting the value for the next plot along. In plot experiments one rarely needs to go beyond first differences and a strategy is recommended of applying first differencing in directions with 8 or more plots. If there are less than 8 plots then fitting an autoregressive process to the undifferenced data is considered to be the best option.

Comparison of neighbour and block design-and-analysis methods suggest (Kempton *et al.* 1994) that when there are strong field patterns then neighbour analysis is clearly more efficient than block methods. Efficiency in this context is measured as the average variance of variety differences from a complete block analysis expressed as a ratio of the average variance from the more complex method. Where trends across trials are small there are indications (Kempton *et al.* 1994) that neighbour analysis based on first differences can result in a potential loss of treatment information and may produce a less efficient analysis than a block analysis.

As mentioned previously the choice of optimal neighbour model involves an exploratory process. For most plant breeders faced with the routine analysis of many trials and measurements this is not a trivial task. In such circumstances the best strategy would seem to be to identify a common neighbour model which fits the majority of cases and apply this to all trials.

Design of series of trials

A well-conducted variety trial with two replicates might be expected to provide estimates of between-variety yield differences with a standard error of approximately 5%. This standard error represents a measure of how similar will be the relative performance of varieties if a trial is repeated at the same place under the same conditions of management, weather, pests and diseases. No matter how accurate it is, a single variety trial is of limited value in predicting the performance of varieties when grown in other locations and seasons. Under Northern European conditions a typical cereal variety trial estimates how two varieties will yield regionally with a standard error of difference of approximately 12%. Thus if a variety yields 110% of another in a trial we can only claim with any confidence that the mean performance of the variety regionally will be somewhere between 100% and 120%.

In spite of large between-trial variation it is nevertheless possible to identify good varieties by combining information from several trials. The cost of an additional trial is not negligible but the loss from failing to identify a good variety is likely to be far more substantial. Here we consider some of the issues involved in determining the optimum allocation of resources across trials.

Sources of G × E variation

The major environmental factors affecting relative yield performance in variety trials may be classified under the heading of centre (location) and year (season). The variety × centre component results from relative performance of varieties changing from centre-to-centre in a way that is similar each year. The variety × year variance arises from differences in variety performance between seasons which are apparent at all centres. The variety × centre × year term represents variety differences which change from centre-to-centre to an extent that is dependent on the season, or variety differences which are affected by seasonal changes at some centres more than at others.

The separation of variety × environment variation into its main components has been done in a number of studies of national crop variety testing programmes, *e.g.*, Talbot (1984) in the UK, Laidig & Utz (1992) in Germany. While there are some differences between crops in the effects of location and season on variety performance, nevertheless a general pattern is apparent. The variety × centre × year term is the dominant component representing approximately two-thirds of total variety × environment variation in the UK data. Some 20% of the remainder is attributable to differences between seasons in the relative performance of varieties and an average of 10% is due to adaptation of varieties to conditions at some centres more than others.

The extent to which this pattern applies in other European countries is illustrated in Table 1 where components of cereal variety yield variation for the UK, Germany and Spain are presented. Although growing conditions might be expected to be quite different between these countries nevertheless there are clear similarities in the response of varieties to environments in those countries.

Table 1. Barley and wheat variety yield variation in some EU countries. Components of yield variation, SD as % of mean yield

	UK	Germany	Spain
Variety × centre	2.2	2.4	2.8
Variety × year	3.1	2.5	4.2
Variety × centre × year	5.6	5.2	5.5
No. of years	5-12	8	4.5
No. of centre	20-26	15-17	20-27

Objectives and criteria

The statistical contribution to the planning of experiments aims at obtaining maximum information with minimum effort. In variety testing we seek information to achieve two primary objectives: selecting varieties which will improve yields to the greatest extent; and minimising the risk of discarding a good variety and the criteria by which we judge the effectiveness of a trials system must reflect these objectives. Two statistical criteria are considered: acceptance probabilities and potential gain.

Potential gain measures the average difference between all varieties entering trials and those finally selected. Gain is a function of the proportion of varieties accepted as well as the efficiency of the trials system. For a fixed proportion of varieties accepted, then the larger is the gain the more efficient is the trials system.

An acceptance probability is the probability that a variety of unknown performance relative to a standard will be accepted. The acceptance probability is influenced by the accuracy of the trials systems, as well as, for any individual variety, the size of the difference between its true performance and the acceptance standard.

Potential gain

Numerical integration methods (Robinson 1984) have been used to produce estimates of the average potential gain to be achieved in trials systems with varying levels of precision and selection rates (Table 2). The calculation of gain assumes that variances are normally distributed and that the correlation between measurement errors in different years, is small, as occurs in UK variety trials.

Estimates in Table 2 are for several ranges of variety performance; since, clearly, it is easier to identify good varieties when the differences between submitted varieties are large rather than small. The range is calculated as the yield of the best variety minus the yield of the poorest variety out of twenty varieties submitted. In the UK the range in average yields of varieties submitted for testing is of the order of 10%.

The advantages of selective discarding of varieties may be seen in Table 2. After the first year of testing as many as 60% of the poorer varieties can be removed without noticeable reductions in average gain; as a result, considerable savings could be made in the total numbers of plots sown.

While there are advantages in selectively discarding varieties it is important not to discard too many. There is a noticeable reduction in percentage average gain by rejecting as many as 80% of varieties after one year of trials. A rough guide to optimal selection

Table 2. Effect on gain of changes in discarding rate, trial precision and variety differences

Range of variety differences (%)	SE of variety mean (%)	Total % varieties discarded by end of year							
		Year 1	90	-	60	80	-	60	80
		2		90	90	90	-	80	-
		3					90	90	90
		% average potential gain							
5	1		2.4	2.7	2.7	2.7	2.8	2.8	2.8
	2		1.6	2.1	2.0	2.0	2.3	2.2	2.1
	3		1.2	1.6	1.6	1.5	1.8	1.8	1.6
	4		0.9	1.2	1.2	1.2	1.5	1.4	1.3
	5		0.7	1.0	1.0	1.0	1.2	1.2	1.1
10	1		4.4	4.5	4.5	4.5	4.6	4.6	4.6
	2		3.6	4.1	4.1	4.0	4.2	4.2	4.2
	3		3.0	3.6	3.5	3.4	3.8	3.8	3.6
	4		2.4	3.1	3.1	2.9	3.4	3.4	3.2
	5		2.1	2.7	2.7	2.5	3.0	3.0	2.8
20	1		9.2	9.3	9.3	9.3	9.3	9.3	9.3
	2		8.7	9.0	9.0	9.0	9.2	9.2	9.1
	3		8.0	8.6	8.6	8.6	8.9	8.9	8.8
	4		7.3	8.1	8.1	8.0	8.5	8.5	8.3
	5		6.6	7.6	7.6	7.5	8.1	8.0	7.8

rates is to discard the same proportion at every stage. Thus in a two-stage system, in which we finally select 10 varieties out of 100, approximately two-thirds of the candidate varieties might be discarded at the end of the first stage.

Acceptance probabilities

The consequences of changes in trials systems on the probability of a variety being accepted may also be studied by computer simulation. Table 3 shows the risk of a good variety being rejected, *i.e.*, one minus the probability of acceptance, because it is not ranked in the top third of varieties by its trial results. In general, varieties are likely to be accepted as showing a clear improvement in a character if their performance in trials indicates a ranking in the top third of varieties. At the other end of the scale, varieties are likely to be rejected whose performance places them in the bottom third of varieties, unless there are compensating characters.

It will be seen from Table 3 that as the accuracy of trials decreases then the probability of rejection increases. A corollary is that the probability of wrongly accepting varieties that do not meet the standards will also increase.

From Table 3, if the range of variety yields is 10% and the standard error of mean if 4%, then the best variety in 100 has a 5% chance of not being ranked in the top third of varieties after two years of trials. Under similar circumstances the fifth-best variety has a

Table 3. Risk of a good variety producing yields in trials which rank it in the lower two-thirds of varieties after two years of trials

Range of variety differences (%)	SE of variety mean (%)	True rank of variety (out of 100)					
		1	5	10	1	5	10
		% risk of variety being ranked in lower two-thirds			% risk of variety being ranked in lowest third		
5	1	0	1	3	0	0	0
	2	5	12	19	0	1	2
	3	16	25	31	2	5	7
	4	25	34	39	5	9	11
	5	32	40	44	8	12	14
10	1	0	0	0	0	0	0
	2	0	1	3	0	0	0
	3	1	6	11	0	0	1
	4	5	12	19	0	1	2
	5	10	10	26	1	3	4
20	1	0	0	0	0	0	0
	2	0	0	0	0	0	0
	3	0	0	1	0	0	0
	4	0	1	3	0	0	0
	5	0	3	7	0	0	0

1% chance of being ranked in the bottom third of varieties.

Application of the criteria

Tables 2 and 3 can be applied generally in many testing circumstances. To illustrate, we use estimates of variety yield variability to examine the effects of changes in a trials system.

Reducing the number of testing years from two to one has the greatest effect of any change that might be considered. In most practical circumstances there is a 25% loss in gain as a result of changing the testing period from two years to one year.

Increasing the number of trials per year beyond 12 or 13, which gives a standard error of 3.6%, has only a small effect on gain. Of course, more extensive trialling may be needing if regional or other special requirements have to be satisfied.

In a two year trial system with eight trials per year there is a 1 in 20 chance of the best variety in 100 being identified as mediocre, *i.e.*, with a trial performance that ranks it in the lower two-thirds of varieties. There is a 1 in 50 chance of the tenth best variety being ranked in the lowest third of varieties. In many situations such risks might be considered as unacceptable. The risk can only be lessened by increasing the number of trials per year or the number of trial years.

The allocation of testing effort between years is important. Additional trials at later stages cannot compensate for insufficient trials in earlier years (Table 4). The good varieties that are discarded on the basis of poor early results do not get a chance to show their true performance later. Only when the number of varieties taken into the later stages is very large there is any compensation. On this evidence, the UK system, with 10-13 initial trials testing large numbers of varieties followed by more extensive trialling with few varieties, provides a reasonably sound basis for variety improvement.

An increase in within-trial replication from 2 to 3 plots has relatively little effect on precision. On average, standard errors for means based on a single trial (Table 5) are

Table 4. Potential gain when selecting the best 5% of varieties in a 5-year trials system with limited numbers of trials in early years*

Number of trials per year in years		Average percentage gain when selecting best 5% but discarding after second year the proportion		
1-2	3-5	66%	80%	90%
13	50	1.78	1.76	1.68
13	13	1.72	1.71	1.65
6	50	1.74	1.71	1.50
3	50	1.70	1.64	1.49

* Example from UK variety yield trials with variability estimates as in Table 1

reduced from 8.4% to 7.8% in the UK situation; for means based on 13 trials the effect is not noticeable that is quoted.

Discussion

In the application of the criteria to the choice of a trials system, gain and acceptance probability cannot be regarded as alternatives. They each describe different aspects of the same trials system. One deals with risk to the breeder. The other is concerned with gains to the country. Both aspects are important and both must be taken into account.

The criteria described have limitations. Potential gain does not measure what might be achieved in commercial agriculture. It only indicates what should happen if all selected varieties are grown to the same extent. In practice, the best varieties are more widely grown and these are more likely to be selected whatever trials system operates.

Also, the criteria attempt to predict what may happen in future based on an average of past experience and, if a future season is abnormal, the average may not be a good prediction. Nevertheless, in the absence of information about the future an average provides a reasonable basis for decision.

The key role of varieties in crop improvement makes it important that the procedures for identifying good varieties are as effective as possible. The criteria provide an internal audit of the effectiveness of some of these procedures.

Table 5. Average standard error of variety mean-over-trials dry matter yield in a single year of UK crop variety trials with two replicates per trial

Number of trials/year	SE of variety mean as % of mean yield		
	Mean*	Minimum	Maximum
1	8.4	7.7	9.1
2	6.4	5.8	6.6
3	5.4	4.8	5.7
4	4.8	4.3	5.3
5	4.6	3.9	5.0
6	4.3	3.6	4.8
7	4.2	3.4	4.7
8	4.0	3.2	4.6
9	3.9	3.1	4.5
10	3.8	3.0	4.4
11	3.7	2.9	4.4
12	3.6	2.8	4.3
13	3.6	2.7	4.2
50	3.0	2.0	3.8

* Mean, minimum and maximum standard errors for several crops including grain dry matter of wheat, barley and oats; total annual herbage dry matter production of perennial ryegrass; root dry matter yield of swedes; digestible organic matter yield of kale; dry matter seed yield of oilseed rape

References

Baird, D. & R. Mead, 1991. The empirical efficiency and validity of two neighbour models. Biometrics 47: 1473-1487.

Besag, J. & R. Kempton, 1986. Statistical analysis of field experiments using neighbouring plots. Biometrics 42: 231-251.

Cullis, B.R. & A.C. Gleeson, 1991. Spatial analysis of field experiments - an extension to two dimensions. Biometrics 47: 1449-1460.

Cochran, W.G. & G.M. Cox, 1957. Experimental Designs, 2nd. Edition. Wiley, New York.

Gilmour, A.R., 1992. TwoD: a program to fit a mixed linear model with two dimensional spatial adjustment for local trend. NSW Agricultural Biometric Bulletin No 1: 1-73.

Gleeson, A.C. & B.R. Cullis, 1987. Residual maximum likelihood (REML) estimation of a neighbour model for field experiments. Biometrics 43: 277-288.

Green, P.J., C. Jennison & A.H. Seheult, 1985. Analysis of field experiments by least squares smoothing. Journal of the Royal Statistical Society, Series B 47: 299-315.

Kempton, R.A., J.C. Seraphin & A.M. Sword, 1994. Statistical analysis of two dimensionsal variation in variety yield trials. Journal of Agricultural Science: in press.

Laidig, F. & H.F. Utz, 1992. Combining results from nonorthogonal trial series over-years. Biuletyn Oceny Odmian 24-25: 55-68.

Lill, W.J., A.C. Gleeson & B.R. Cullis, 1988. Relative accuracy of a neighbour method for field trials. Journal of Agricultural Science, Cambridge 111: 339-346.

Martin, R.J., 1990. The use of time-series models and methods in the anlaysis of agricultural field trials. Communications in Statistics - Theory and Methods 19: 55-81.

Nguyen, N-K. & E.R. Williams, 1993. An algorithm for constructing efficient resolvable row-column designs. Australian Journal of Statistics 35: 363-370.

Patterson, H.D. & E.A. Hunter, 1983. The efficiency of incomplete block designs in National List and Recommended List cereal variety trials. Journal of Agricultural Science, Cambridge 101: 427-433.

Patterson, H.D. & E.R. Williams, 1976. A new class of resolvable incomplete block designs. Biometrika 63: 83-92.

Pithimcharumlap, M., K.E. Basford & W.T. Federer, 1993. Neighbour analysis with adjustment for interplot competition. Australian Journal of Statistics 35: 263-270.

Robinson, D.L., 1984. A study of sequential variety selection systems. Journal of Agricultural Science, Cambridge 102: 119-126.

Talbot, M., 1984. Yield variability of crop varieties in the UK. Journal of Agricultural Science, Cambridge 102: 315-321.

Wilkinson, G.N., S.R. Eckert, T.W. Hancock & O. Mayo, 1983. Nearest neighbour (NN) analysis of field experiment (with Discussion). Journal of the Royal Statistical Society B 45: 151-211.

Williams, E.R., 1986. Row and column designs with contiguous replicates. Australian Journal of Statistics 28: 154-163.

William, E.R., 1986. A neighbour model for field experiments. Biometrika 73: 279-287.

Williams, E.R. & M. Talbot, 1993. ALPHA+ Experimental designs for variety trials. Design User Manual. CSIRO, Canberra & SASS Edinburgh.

Relationship between molecular and morphological distances in a maize inbred lines collection. Application for breeders' rights protection

A. Bar-Hen[1] *& A. Charcosset*[2], [1]*G.E.V.E.S. La Minière, 78280 GUYANCOURT cedex, &* [2]*I.N.R.A. station de génétique végétale, ferme du Moulon, 91190 GIF* ^{s}*/ YVETTE, FRANCE*

Abstract

Molecular markers have been proposed as a tool for breeders' right protection (Beckmann & Soller 1983, Smith *et al.* 1991). In order to bring complementary information about their possible use for this purpose, 150 maize inbred lines were described using: (i) RFLP markers; (ii) morphological characters used in present distinctiveness studies. The aim of the experiment was to investigate the relationship between the distances computed for both types of traits.

100 probes were chosen to optimize genome coverage, and three restriction enzymes were used. A total of 222 probe*enzyme combinations were polymorphic and interpretable. For each pair of lines, we calculated the best estimate of the percentage of loci in common.

Evaluating the precision of distance estimation is important for previous and other purposes. Studies have addressed this topic in the general case of populations (Nei & Roychoudhury 1973, Mueller 1979). The jackknife method has been used in the specific case of inbred lines to investigate the effect of loci sampling (Melchinger *et al.* 1991, Bernardo 1993). We demonstrated that the jackknife estimate of the variance is equivalent to the maximum likelihood estimate of the variance of a binomial law. Implications of the analytical estimate of the variance are discussed.

For the morphological characters we used Mahalanobis distance. The relationship between this distance and the distance computed using molecular data has a "triangular" shape. Low molecular distances are always associated with low morphological distances, whereas high molecular distances may be associated with any morphological distances, either low or high. This should be due to the following reasons: (i) different genotypes may lead to the same phenotype (such as ++++−−−− and −−−−++++ if eight homozygous loci are considered); (ii) molecular markers are supposed to be neutral, i.e.

to have no direct effect on the morphological traits under study. Thus, the relationship will depend on the linkage disequilibria between molecular markers and the loci involved in the morphological traits. Consequences for breeder's right protection are discussed.

Introduction

Following UPOV recommendation, protection of a new maize variety includes the study of its distinctiveness with pre-existing varieties. Currently, comparisons are essentially performed on the basis of morphological data. Genetic markers could provide an additional tool for the comparison of varieties, as was discussed by Soller & Beckmann (1983) and Beckmann & Soller (1983). The possible use of genetic markers in a distinctiveness scheme was discussed by Smith *et al.* (1991). It clearly depends on: (i) the precision of marker distance evaluation, to determine the number of markers to be used to get reliable estimates of the marker distance; (ii) the relationship between marker distance and morphological descriptors, to determine their relative roles in the distinctiveness scheme. The aim of this study is to consider elements of these two aspects.

Material and Methods

Germplasm under study

145 maize inbred lines were described using: (i) RFLP markers; (ii) morphological characters used in present distinctiveness studies. These lines were obtained in various breeding programs run by private companies, as well as public institutes, and are adapted to various climatic conditions that can be found in France. They were chosen to be representative of the different groups of Maize germplasm used in Europe. For reasons of confidentiality, the lines were coded and the relatedness of the lines is unknown.

RFLP analyses

For RFLP analysis, 100 probes were chosen to optimize genome coverage, and three restriction enzymes were used. DNA was extracted from a sample of 15 plants according to the method described by Rogers & Bendich (1988). To evaluate the precision of the method, five lines were repeated and therefore were studied two times. Depending on the quality of the result, the markers (probe*enzyme combinations) were noted from *A* (very good quality) to *D* (non interpretable). Table 1 gives the repartition of the probes with

Table 1. First description of the 222 markers

	EcoRI	HindIII	EcoRV	Total
Quality A	11	25	13	49 (22%)
Quality B	35	42	37	114 (51%)
Quality C	15	19	25	59 (27%)
Total	61 (27%)	86 (39%)	75 (34%)	222

respect to the quality and the enzyme used.

A total of 222 probe*enzyme combinations were polymorphic and interpretable. The average distance between two neighbour probes is 13.4 cM. The average distance varies according to the chromosome from 8.6 to 15.4. The average number of levels of migration (band levels) for the 222 interpretable markers is 4.96. It is varying from 4.43 for markers of quality *A* to 5.36 for markers of quality *C*. Markers were classified in four classes:

* unilocus probes where an allele corresponds to exactly one band;
* unilocus probes where an allele can correspond to more than one band;
* multilocus probes where an allele corresponds to exactly one band;
* multilocus probes where an allele can correspond to more than one band.

For 15 markers, it was not possible to conclude about the number of loci.

Quality of the markers is fundamental to obtain reproducible results and reliable distances between varieties. To test the method, five lines were repeated. For one replicated line, no difference was observed between the two replicates. For another replicated line, seven differences were observed. It has to be noted that 6 of these differences were observed on markers of quality *C*. For the three other replicated lines, between one and three differences were observed.

Computation of marker distances

For each pair of lines, we calculated the best estimate of their percentage of loci in common. For this purpose we computed Roger's distance for single locus probe*enzyme combinations and Nei and Li's distance for multilocus probe*enzyme combinations. A synthetic distance was subsequently derived as the average of previous distances, weighted by their respective number of loci.

Morphological description and distance computation

Quantitative traits under study are listed in Table 2. In the years 1989, 1990, 1991 and 1992, the experimentation network involved three locations (La Minière, near Paris, Le

Table 2. Name of the quantitative trait

Length of ear (mm)
Diameter of ear (mm)
Diameter of cob (mm)
Length of plant (cm)
Height of ear (cm)
Number of rows
Width of blade (mm)
Length of main axis above lowest side branch (cm)
Length of main axis above highest side branch (cm)
Date of male flowering (days)

Magneraud, centre-west of France and Saint Martin de Hinx in the South-West of France). Depending on line earliness, traits are evaluated in at least 2 convenient locations. Each location was planted with two replications in a block design. Each plot had twenty plants in a single row. The distance between two plants in the same row was 25 cm, and the distance between two rows was 80 cm. Measures of quantitative traits were taken over the ten most representative plants.

An analysis of variance was performed to remove the environmental effects. Then Mahalanobis distance was computed on the residuals.

Statistical properties of the marker distance

Evaluating the precision of distance estimation is important to determine the number of markers to be used for distinctiveness studies. Studies have addressed this topic in the general case of populations (Nei & Roychoudhury 1973, Mueller 1979). Under the assumption of a random sampling of the loci, the variance of the estimate of the marker distance has been estimated in several studies using the jackknife approach (Melchinger *et al.* 1991, Bernardo 1993). The jackknife estimate of the variance (Efron 1982, formula 1.3) is:

$$\hat{\sigma}^2_{jack} = \frac{N-1}{N} \sum_{l=1}^{N} (\hat{p}_l - \hat{p})^2 \, , \tag{1}$$

where $\hat{p}_l$ is the proportion of identical loci estimated from the sample of loci after excluding the locus L_l. We have proved (Bar-Hen & Charcosset, submitted) that:

$$\hat{\sigma}^2_{jack} = \frac{1}{N-1}\,\hat{p}(1-\hat{p})\ . \tag{2}$$

Thus, the jackknife estimate of the variance of $\hat{p}$ is equivalent, modulo a $N/(N-1)$ factor, to the classical binomial estimate of the variance of a proportion:

$$\hat{\sigma}^2_{bin} = \frac{1}{N}\,\hat{p}(1-\hat{p})\ . \tag{3}$$

The discrepancy between the two estimates ($N/(N-1)$ factor) is related to the fact that one probe is released in the estimation of $\hat{p}_l$, and therefore the sampling variance is estimated for $N-1$ loci. The fact that both approaches lead to the same result is expected since both assume an independent sampling of the loci.

It is important to underline that previous approaches assumed an independent sampling of the loci. No assumptions are necessary concerning the statistical associations between loci. Practically speaking, this means that linkage disequilibrium in the population of analysed genotypes has no effect on the variance of the estimation of the distance. However, it should be noticed that the hypothesis of an independent sampling of the loci is not always fulfilled. This is, for instance, the case of species for which a large number of probes has been developed and mapped (such as maize and tomato). In these situations, selected sets of probes that optimize genome coverage have generally been defined (see for instance the maize core-map, Gardiner *et al.* 1993) and are used for distance estimation. This strategy leads to an increased precision of the estimation of the distance, as discussed by Bar-Hen & Charcosset (submitted).

The computation of estimates using the jackknife method is very time-consuming. If it takes 3 minutes to compute a set of distances based on 180 loci, it will take around nine hours to compute the jackknife estimate of the variance. Thus, the use of formula (3) allows faster computations. Distributional results have another advantage: it is possible to derive confidence intervals for the estimates (see Table 3) and tests for the comparison of two distances.

Let p_{ij} be the marker distance between lines i and j and let $\hat{p}_{ij}$ be the estimate of p_{ij} computed with N independent observations. Two cases have to be considered:

* the first situation of interest is when the two distances are computed from different lines. In this case the two binomial tests are independent. It has been proved that, asymptotically:

Table 3. Limits of the confidence intervals (5% level) for the proportion of identical (or different) loci, $\hat{p}$ estimated with a random sample of N marker loci

N	$\hat{p}$				
	0%	10%	20%	30%	50%
10	0-45.0	2.5-44.5	6.6-55.6	12.1-65.2	26.2-81.2
20	0-16.8	3.2-31.7	8.6-43.6	15.4-54.3	31.5-72.8
30	0-11.5	3.8-26.5	9.9-38.5	17.3-49.4	34.3-68.7
50	0-7.1	4.5-21.8	11.5-33.7	19.5-44.5	37.4-64.4
100	0-3.6	5.6-17.6	13.5-29.2	22.1-40.0	40.8-60.1
200	0-1.8	6.6-15.0	15.1-26.2	24.2-36.9	43.3-57.1

$$N\frac{\hat{p}_{ab} - \hat{p}_{cd}}{\sqrt{\hat{p}_{ab}(1 - \hat{p}_{ab}) + \hat{p}_{cd}(1 - \hat{p}_{cd})}} \sim N(0,1) \quad \text{as } N \to +\infty \ ;$$

* the second question of interest is to determine if line A is closer to line B than to line C. This is, for instance, important when assigning inbred lines to heterotic groups. This means that we want to test:

$$H_0: \ d(A,B) = d(A,C) \ .$$

In this case the two binomial laws are not independent. It has been proved that, asymptotically:

$$N\frac{\hat{p}_{ab} - \hat{p}_{ac}}{\sqrt{\hat{p}_{ab} + \hat{p}_{ac} + 2\hat{p}_{abc} + (\hat{p}_{ab} - \hat{p}_{ac})^2}} \sim N(0,1) \quad \text{as } N \to \infty \ .$$

Experimental results

Description of the polymorphism at marker loci

The 222 markers represent 1546 profiles. (442 for EcoRI, 570 for HindIII and 534 EcoRV). Therefore, there are on average of seven profiles per probe*enzyme combination.

421 (specific) profiles are present in only one line. 106 lines have at least one specific profile. Among these lines, one line (line 35) has 78 specific profiles. It is difficult to conclude if this line is very original or if there was a problem during the experiment.

The case of line 50 is also of interest: its most characteristic profile is also present in 13 other lines. This line may be considered as a founder line, or could be a combination of common profiles.

Redundancy of information of the enzymes for a given probe

For each probe and each couple of enzymes, we have computed the number of lines which have the profile i with the enzyme A and the profile j with the enzyme B. There exists many possible ways to compute a coefficient of association in this kind of table of contingency. Since the number of lines is constant (145) and the number of profiles is variable (and so the number of rows and columns in the contingency table) we computed the Cramer's V coefficient of association (see Bishop *et al.* 1975). The correlation between EcoRI and HindIII was studied with 54 probes; the correlation between EcoRI and EcoRV was studied with 45 probes and the correlation between HindIII and EcoRV with 68 probes. The association is always high and positive (except for the probe UMC132 with the enzymes EcoRI and HindIII). To quantify the magnitude of these associations, we used Cramer's V coefficient of association. The theoretical range of V is between -1 and 1. For the three pairs of enzymes, the median of the V is always between 0.7 and 0.8. The discrepancy around these values is comparable for the three pairs of enzymes.

This suggests that, in general, the use of a single enzyme per probe should be recommended, since a second enzyme brings little additional information.

Relationship between marker distance and morphological distance

Figure 1 clearly shows that the relationship between marker distance and morphological distance is not linear but displays a "triangular" shape. Low marker distances are systematically associated with low morphological distances. On the other hand, high marker distances are associated either to high or low morphological distances. Thus, marker divergence behaves as a limiting factor of morphological divergence. Two explanations can be proposed for this relationship.

1. First of all, it is clear from quantitative genetics theory that two different combinations of genes may lead to the same phenotype. We will assume that a trait is controlled by eight biallelic loci with equal effects and no epistasis, quoting + and $-$ the favourable and unfavourable homozygous genotypes, respectively. It is then clear that genotypes $----++++$ and $++++----$ have the same phenotypic value, although they are different at every individual locus. This generates a triangular relationship between the distance for a quantitative trait and the proportion of QTLs (involved in the

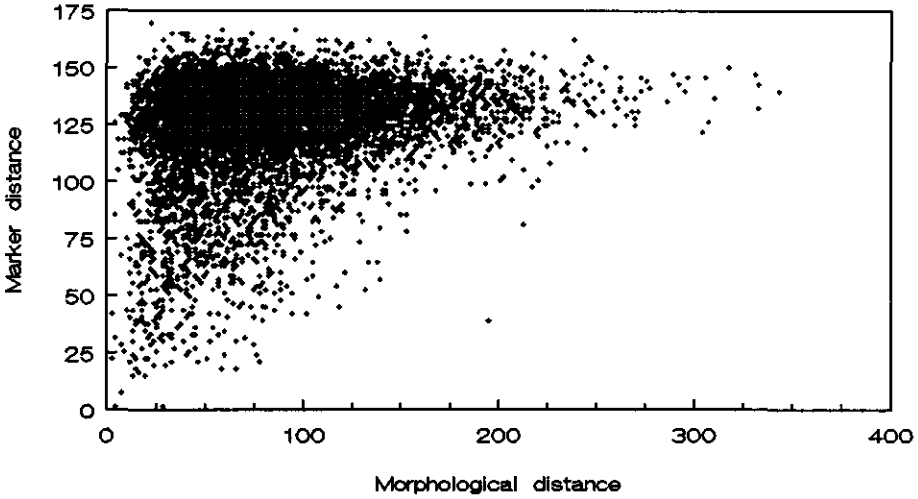


Figure 1. Marker distance versus morphological distance

variation of this trait) for which two lines differ, as was illustrated by Charcosset (1992). Similar relationships would be obtained for distances computed from several quantitative traits (which is the case of the Mahalanobis distance), provided each of them is controlled by several loci.

2. Since it is not possible to estimate directly the proportion of QTLs for which to lines differ, distance is computed using genetic markers. Linkage disequilibrium between markers and QTLs will affect the relationship between morphological distance and marker distance. If there is linkage equilibrium between markers and QTLs, the two distances will vary independently; high and low marker distance will correspond to similar morphological distance.

Properties of the relationship between distance at marker loci and the proportion of QTLs for which the two lines display different alleles are similar to those described by Charcosset & Essioux (1994) for the relationship between marker distance and heterosis. Strong relationships (in the sense of a large correlation coefficient) are expected (i) for couples of lines which are related by pedigree or (and) belong to a same genetic group, (ii) for couples of lines which represent a mixture of within and between groups comparisons. On the other hand, no relationship is expected for the between groups couples. Experimental results by Smith *et al.* (1991) and Melchinger *et al.* (1992) clearly illustrate the result of these tendencies: when considering the total range of couples the relationship appears very strong; when considering only rather small marker distances the relationship remains very strong, however when considering only rather large distances, the relationship gets very weak. Properties of (1) and (2) both contribute to the general triangular tendency of Figure 1.

Discussion and conclusion

Table 3 illustrates that the determination of the number of marker loci to be introduced in distance computation depends on the cost of the experiments and the objective of the study. Approximately, using 50 loci is efficient to determine whether the similarity between two lines is 10% or 30%, 30% or 50%, 50% or 70%. Such a precision should allow to attribute lines to heterotic groups, provided there is some relatedness within the groups. On the other hand, if two lines differ for 7.1% of the loci, they will be declared identical for 2.5% of the experiments, which may be considered as a too important type 2 error level for distinctiveness studies.

Experimental results concerning the relationship between marker and morphological distances illustrate that the use of morphological traits to protect germplasm is efficient: if two lines differ at the morphological level, they will differ at marker level. Thus, the probability that a line is registered, although it is very close (at the DNA level) to a pre-existing line, appears very low. However, it is clear that markers bring complementary information in case of morphological similarity. They would allow to discriminate between: (i) a close similarity which is the result of independent breeding efforts (different combinations of genes), (ii) close similarity due to copies or very high relatedness.

Since the relationship between marker distance and morphological distance is not linear, a sequential use of both types of information (Smith *et al.* 1991) should be preferable to a linear combination of the two distances. Further investigations are currently carried out on this subject.

Acknowledgements

This research was partially funded by the French Ministry of Agriculture and SEPROMA association. We are grateful to all the persons who obtained the data used in this study, and specially to A. Murigneux for RFLP data and D. Guérin for field observations. We are also grateful to all the participants to the "distinction working group" for helpful

discussions.

References

Bar-Hen, A. & A. Charcosset. Precision of the estimation of genetic distance between inbred lines using molecular markers. Submitted.

Beckmann, J.S. & M. Soller, 1983. Restriction fragment length polymorphisms in varietal identification and genetic improvement: methodologies, mapping and costs. Theor Appl Genet 67: 35-43.

Bernardo, R., 1993. Estimation of coefficient of coancestry using molecular markers in maize. Theor Appl Genet 85: 1055-1062.

Bishop, Y.M.M., S.E. Fienberg & P.W. Holland, 1975. Discrete Multivariate Analysis: Theory and practice. Cambridge, MA, The MIT Press.

Charcosset, A., 1992. Prediction of heterosis. Reproductive Biology and Plant Breeding. Y. Dattée, C. Dumas & A. Gallais (editors). Springer-Verlag. p. 355-369.

Charcosset A. & L. Essioux, 1994. The effect of heterosis on the relationship between heterosis and heterozygosity at marker loci. Theor Appl Genet: in press.

Efron, B., 1982. The bootstrap, the jackknife, and other resampling plans. S.I.A.M., Philadelphia.

Gardiner, J.M., E.H. Coe, S. Melia-Hancock, D.A. Hoisington & S. Chao, 1993. Development of a core RFLP map in maize using an immortalized F2 Population. Genetics 134: 917-930.

Melchinger, A.E., M.M. Messmer, M. Lee, W.L. Woodman & K.R. Lamkey, 1991. Diversity and relationships among U.S. maize inbreds revealed by Restriction Fragment Length Polymorphisms. Crop Sci. 31: 669-678.

Mueller, L.D., 1979. A comparison of two methods for making statistical inferences on Nei's measure of genetic distance. Biometrics 35: 757-763.

Nei, M. & A.K. Roychoudhury, 1973. Sampling variances of heterozygosity and genetic distance. Genetics 76: 379-390.

Rogers, O.S. & A.J. Bendich, 1988. Extraction of DNA from plant tissues. Plant Molecular Biology Manual A6: 1-10

Smith, J.S.C., O.S. Smith, S.L. Bowen, R.A. Tenborg & S.J. Wall, 1991. The description and assessment of distances between lines of Maize. III A revised scheme for the testing of distinctiveness between inbred lines utilizing DNA RFLPs. Maydica 36: 213-226.

Soller, M. & J.S. Beckmann, 1983. Genetic polymorphism in varietal identification and genetic improvement. Theor Appl Genet 47: 179-190.

Theoretical investigations into the use of subpopulations in recurrent selection of sugar beet

D.C. Borchardt & H.H. Geiger, University of Hohenheim, Institute of Plant Breeding, Seed Science, and Population Genetics (350), D-70593 Stuttgart, Germany

Abstract

Population improvement by means of recurrent selection can be carried out using a single main population, or by using several subpopulations staggered by one year. To compare these two approaches theoretically, several methods of recurrent testcross selection in diploid sugar beet (*Beta vulgaris* L.) have been optimized with respect to the number of test units to be evaluated, the number of test locations, the selection intensity, and the number of recombination units. Additionally, restrictions regarding the effective population size and available financial capacities were imposed. Genetic gain per year was calculated for recurrent selection and line development for recoverable sugar yield as optimization criterion. S_1-line testcross selection with a four-year recurrent selection cycle and combined S_1-line/S_3-line testcross selection with a six-year recurrent selection cycle were compared in this study. Subpopulations were assumed to be interlocked during recombination. The best way of interlocking subpopulations was to include only genotypes from the next older subpopulation in a 6 : 1 ratio. Using subpopulations instead of a single main population caused a slight reduction in gain of 3 - 11% for recurrent selection and 0 - 7% for line development. Practical advantages of using subpopulations are discussed.

Introduction

In this paper we present results of theoretical studies on the usefulness of subpopulations in recurrent selection (RS) of sugar beet (*Beta vulgaris* L.). The aim of RS is to increase the frequency of favourable alleles in a population with a minimum reduction of genetic variability by keeping the effective population size sufficiently large (Hallauer & Miranda 1988). An overview of the general scheme of RS, adapted from Strahwald & Geiger (1988), is given in Figure 1. The population to be improved is subdivided into so-called selection units (SU). For these SU a selection decision has to be made based on the performance of test units (TU). Recombination units (RU) of the selected fraction are intercrossed to build up the improved population for the next RS cycle. Goal units (GU) for RS are randomly chosen plants from the improved population.

The goal of hybrid breeding is the development of inbred lines to produce experimental hybrids and varieties. In every RS cycle a new "one way" selection for line development (LD) is started. This extension is indicated in Figure 1 with dashed lines. Goal units for LD are the best SU or lines derived thereof.

The genetic units (SU, TU, RU, and GU) may be partially or completely identical depending on the RS or LD method considered.

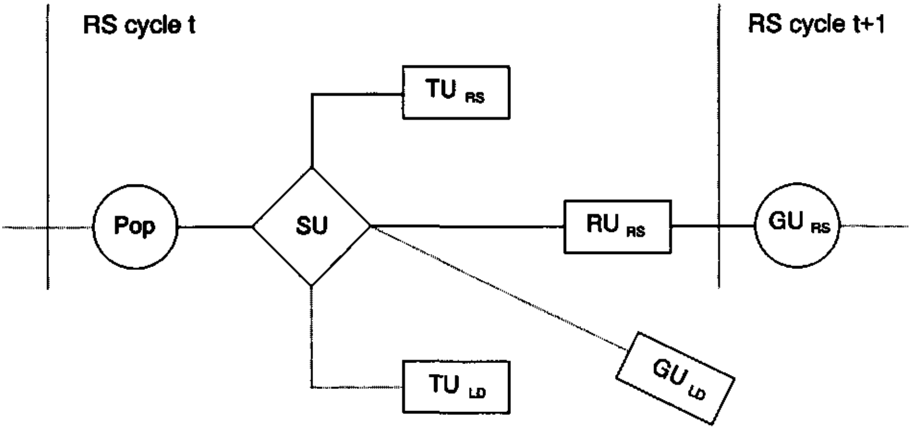


Figure 1. General scheme of recurrent selection (RS) and line development (LD) (Pop = population, SU = selection units, TU = test units, RU = recombination units, GU = goal units)

There are two principal alternatives in dealing with RS populations. Figure 2 gives an example for the improvement of a single main population with cycle length $Y_{RS} = 4$. Starting from the initial population, testcrosses to be used as TU are produced in the second year. In the third year these TU are evaluated in a field trial. The selected fraction is recombined in the fourth and last year of the RS cycle to form the new population. In this approach all breeding steps occur once per cycle. After the evaluation phase the best lines can be used for LD, also once per cycle. The budget per cycle is equal to r times the yearly budget, r being the RS-cycle length.

Figure 2. Improvement of a single main population with RS-cycle length $Y_{RS} = 4$; → LD = use of selected fraction for line development

Cycle	Year	RS	
t	1	Population	
	2	Testcrossing	
	3	Evaluation	→ LD
	4	Recombination	
t + 1	5	Population '	

As an alternative, the breeder may divide the base population into r subpopulations and stagger their RS cycles by one year (Figure 3). With such an approach, each breeding step (Pop, TC, EV and Rec) occurs in successive years within subpopulations but in each year simultaneously across subpopulations. The budget per cycle of a subpopulation then equals the budget per year for the total system.

Year	Sp I		Sp II		Sp III		Sp IV	
1	Pop		.		.		.	
2	TC		Pop		.		.	
3	EV	→ LD	TC		Pop		.	
4	Rec		EV	→ LD	TC		Pop	
5	Pop'		Rec		EV	→ LD	TC	
6	.		Pop'		Rec		EV	→ LD
7	.		.		Pop'		Rec	

Figure 3. Improvement of $r = 4$ staggered subpopulations (Sp I-IV) (Pop = population, TC = testcrossing, EV = evaluation, Rec = recombination, Pop' = improved population; → LD = use of selected fraction for line development)

Gene flow between subpopulations can be achieved by recombining genotypes not only within a given subpopulation but also with genotypes from previously completed cycles of other subpopulations. This interlocking of subpopulations should limit gene loss and genetic differentiation between subpopulations due to random drift.

So far, this subpopulation approach has not been considered in selection theory. To compare the main-population with the subpopulation approach we calculated the expected genetic gain for both alternatives of various optimized breeding schemes.

Description of the model

Methods of recurrent testcross selection in diploid sugar beet have been optimized regarding the number of TU to be evaluated, the number of test locations, the selection intensity, and the number of RU. Several assumptions had to be made for these optimizations.

Selection was aimed at improving general combining ability for recoverable sugar yield. The efficiency of selection methods was judged by the expected selection response per year, averaged over RS and LD.

Estimates for quantitative genetic and economical parameters were derived from official trials and from trials of a private sugar beet breeding company (KWS AG, Einbeck, Germany). The error variance was assumed to be equal to the genotypic variance of a population ($\sigma_e^2 = \sigma_t^2 = 24$ qt^2 ha^{-2}). Ignoring epistasis the total genotypic variance was divided into an additive and a dominance variance according to a 5 : 1 ratio. The ratio of genotypic, genotype by location, genotype by year, and genotype by location by year interaction variance was assumed to be 1 : 0.2 : 0.2 : 0.4. Costs of single breeding activities are described in detail by Borchardt (1994). The financial capacities were fixed at 1 million DM.

Every year a certain number of genotypes has to be provided for development of experimental varieties (N_{LD} per year = 10).

As a measurement for the reduction of genetic variability during recombination the yearly inbreeding rate was chosen to be below a constant limit of 1%. The inbreeding rate was calculated as the probability of crossing relatives multiplied by the coefficient of coancestry. Recombination units were assumed to be intercrossed in a complete factorial design and selection effects during the choice of RU were neglected.

Genetic gain per year for subpopulation II interlocked with subpopulation I was calculated as:

$$G = \varrho\, \sigma_y \left[\frac{i_{\mathrm{II}}}{Y}\, z_{\mathrm{II}} + \frac{i_{\mathrm{I}}}{Y+1}\, z_{\mathrm{I}} \right] ,$$

where G is the genetic gain per year, averaged over RS and LD, in qt ha^{-1}, ρ is the correlation between phenotypic value of the TU and genotypic value of the GU, σ_y is the standard deviation of the genotypic value of the GU, i is the selection intensity, Y the number of years, and z is the genetical proportion of each subpopulation that contributes to recombination.

The selection intensity is larger in the older subpopulation I, because fewer genotypes are interlocked than RU taken from subpopulation II. The cycle length increases because the RS cycle of subpopulation I has been completed one year before that of subpopulation II.

Selection schemes

An example for an optimized selection scheme is presented in Figure 4. Selection scheme A is based on a simple S_1L-testcross selection for RS and LD, respectively. After selfing S_0 plants testcrosses with S_1L are produced in the second year. Testcrosses are evaluated in a field trial in the third year. Parallel to testcrossing S_1L are multiplied in isolation cabins and in the third year S_1L^2 stecklings are produced. Based on results from the field trial plants derived from the best stecklings are recombined in the fourth year. In addition, the best lines can be used directly for the production of experimental hybrids.

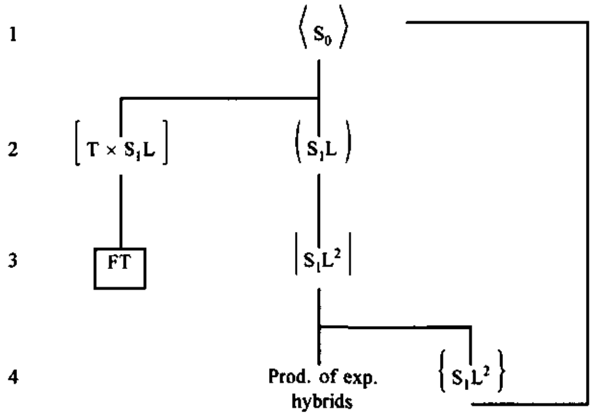


Figure 4. Selection scheme A (FT = Field trial, S_0 = S_0 plant, S_1L = S_1 line, T = tester, S_1L^2 = in isolation multiplied plants of a S_1L). Explanation of symbols: ⟨ ⟩ : selfing, [ ] : production of testcrosses, □ : field trial, { } : recombination, ( ) : multiplication in isolation cabins, | | : production of stecklings

Results and discussion

The opposite effects of cycle length and selection intensity lead to an optimum number

Table 1. Allocations and genetic gains of optimized selection scheme A (N_{TU} , N_{RS} , N_{LD} , P: number of TU, RU, selected lines for LD, and locations, respectively, G_{RS} , G_{LD} , G: genetic gain for RS, LD, and mean of G_{RS} and G_{LD} , respectively)

	Main population	Subpopulation
N_{TU}	877	219
N_{RS}	14	4 + 1
N_{LD}	40	10
P	10	10
G_{RS}	0.528	0.514 (97%)
G_{LD}	1.183	1.183 (100%)
G	0.856	0.848 (99%)

and amount of interlocking of older subpopulations during recombination. As a result of optimization, the best way of interlocking was to include only genotypes from the next older subpopulation in a 6 : 1 ratio.

Table 1 shows the results of optimization for selection scheme A. Using one main population 877 TU are evaluated at ten locations. The best 14 genotypes are selected for RS and the best 40 genotypes for LD. The number of replications in the test was fixed at two although the optimum would be one replication according to the formula for heritability.

With the subpopulation approach 219 TU are evaluated at ten locations with two replications in each of the four subpopulations. Four genotypes from the actual subpopulation, interlocked with the best genotype from the next older one, are recombined and ten genotypes are selected for LD. Genetic gain (G) as optimization criterion is 0.848 qt ha^{-1} and nearly as high as for the main-population approach. Gain for RS is 0.514 qt ha^{-1} (97% of G_{RS} from the main-population approach). Gain for LD is the same as in the main-population approach.

The number of RU using one main population is smaller than the number of RU of all four subpopulations. However, the relative amount of budget for recombination is

Table 2. Comparison of the main-population and subpopulation approach for different selection schemes (S_1L, S_3L, TC, G_{RS} , G_{LD} : S_1 line, S_3 line, testcross, genetic gain for RS and LD, respectively)

Selection scheme	Test units	RS-cycle length	G †	Subpopulation rel. to main population (%) for	
				G_{RS}	G_{LD}
A	S_1L TC	4	0.848	97	100
B	S_1L TC and S_3L TC	6	0.941	89	93

† Genetic gain expected from the subpopulation approach [G = 0.5 (G_{RS} + G_{LD})]

smaller than 1% in both cases.

Table 2 compares genetic gain of the different approaches for two of the calculated selection schemes. Scheme A with RS-cycle length of four years has been shown in Table 1. Selection scheme B is a two-stage selection based on S_1L-testcross performance on the first and S_3L-testcross performance on the second stage. Cycle length of RS is six years. Genetic gain is 11% higher than gain of scheme A. Gain of subpopulation approach is only 89% for RS and 93% for LD.

Thus, the genetic gain of the subpopulation approach is equal to or smaller than the gain of the main-population approach. The reduction of gain is larger with the longer RS cycle of scheme B. However, reduction of genetic gain is not strongly correlated with the cycle length of RS if other selection schemes are considered.

In spite of the lower genetic gain there are important practical advantages in using subpopulations:

The continuity of a breeding program is greater with staggered subpopulations. Genetic gain can be realized every year in one of the subpopulations compared to the only theoretical value of gain per year in the main-population approach. Also there is a continuous output of improved material for LD, when using subpopulations.

Further, progress from selection will show less fluctuation, resulting from genotype by environment interactions, because every year a certain proportion of the breeding material is tested in the field.

Breeding logistics are facilitated by the subpopulation approach because of equal demand of breeding resources, whereas the main population approach leads to a yearly changing requirement of input. When using one main population this disadvantage could be reduced by staggering selection cycles of different breeding programs.

In conclusion, the above mentioned advantages of the subpopulation approach outweigh the slightly reduced expected genetic gain per year. To minimize the latter disadvantage the base population could be split into as few subpopulations as possible with the restriction of having yearly field trials.

Acknowledgement
The authors wish to thank the KWS Kleinwanzlebener Saatzucht AG for providing data and financial support.

References

Borchardt, D.C., 1994. Optimierung von Methoden zur Züchtung von Zuckerrüben. Dissertation. Univ. Hohenheim.

Hallauer, A.R. & J.B. Miranda, 1988. Quantitative genetics in maize breeding. Iowa State Univ. Press, Ames, Iowa.

Strahwald, J.F. & H.H. Geiger, 1988. Theoretical studies on the usefulness of doubled haploids for improving the efficiency of recurrent selection in spring barley. *In:* Proc. 7th Meet. Eucarpia Sect. Biometrics in Plant Breeding. Ås, Norway.

Investigation into the effect of genetic background on QTL expression using three connected maize recombinant inbred lines (RIL) populations

Alain Charcosset, Mathilde Causse, Laurence Moreau & André Gallais,
INRA-UPS-INAPG-CNRS, Station de génétique végétale, Ferme du Moulon, 91190 Gif sur Yvette, France

Key words
quantitative trait locus (QTL), epistasis, recombinant inbred lines (RILs)

Abstract

Three connected RIL populations, derived from three crosses between three parental lines by single seed descent, were analyzed using RFLPs. QTL mapping was done for earliness. The results appeared to be very dependent on the population. The number of QTLs detected in a RIL population appeared to be positively related to the difference in earliness between original parent lines. QTLs may not be detected due to lack of power of statistical tests for QTL detection, allelic relationships between parents and epistatic effects. Data from connected populations make it possible to investigate allelic relationships between parents, *i.e.*, to compare the effects of different genotypes at a given locus.

The genetic design used in this study also allows investigation of epistatic effects, *i.e.*, interactions between a given QTL and other QTLs. Analysis of these interactions indicate that epistatic effects play a role in differences between populations with regard to the presence of QTLs .

Introduction

By using genetic markers, the variation of a quantitative trait within a population can be explained in terms of the map positions of the loci involved (QTLs) and corresponding allele substitution effects. In addition to increasing genetic knowledge, QTL mapping promises successful applications in plant breeding. However, information about QTLs may depend on the environmental conditions, and on the populations which have been used for QTL mapping. Knowledge of the effects of environmental conditions and germplasm is both of fundamental and practical importance.

The influence of genetic background on the positions and effects of QTLs has been illustrated in several studies. Beavis *et al.* (1991) reported results on plant height for four maize populations and noted that no QTL could be consistently identified in all four populations. Of 14 detected QTLs, only two were detected in two different populations. Two of the populations had a common parent. Of the nine QTLs detected in these two populations, only one was detected in both.

Several factors may be responsible for the discrepancies observed between populations with regard to the presence of QTLs. Discrepancies may be due to the power of QTL mapping experiments, especially if QTL effects are small. Other explanations are directly related to the genetics of the trait of interest (Beavis *et al.* 1991).

One explanation concerns allelic relationships between the parents at the QTL. A given QTL may be polymorphic in one population (*i.e.*, the two parents carry different alleles), but monomorphic in another population (*i.e.*, the two parents carry the same allele). Moreover, more than two alleles may be present. As a consequence, the effect of a QTL on the variation of a trait may be very small (or even absent) in one population and large in another population. In the first case, the QTL will not be detected, unless extremely large number of individuals is observed, whereas it will easily be detected in the second case.

A second explanation concerns interactions between the QTL of interest and other QTLs (epistasis). This phenomenon will be referred to as the effect of genetic background on QTL expression.

In order to investigate the importance of these explanations for discrepancies which can be observed between populations, we analyzed the position of QTLs for earliness in maize using three recombinant inbred lines (RIL) populations. These populations have been derived from crosses between three different pure lines, so that every two populations share a common ancestor. This enables us to perform a simultaneous analysis of the three populations together as well as three single population analyses. We will discuss to which extent this type of analysis makes it possible to distinguish between allelic relationships at a QTL of interest, and the effect of genetic background on QTL expression.

Materials and methods

Plant material

The experimental material has been derived from the three possible F_1s between three maize inbred lines:

- an early flint line of European origin (*a*);

- an early dent line of US origin, of which the earliness is similar to that of the early flint line (*b*);

- a dent line from the Iodent group, *i.e.*, from a distinct US origin, which, in the north of France, is approx. 15 days later with regard to silking time than the two previous lines (*c*).

The three hybrids showed large heterotic effects for yield. Crosses between lines of the Iodent group and the European flint line are known to display superior yield potential in the Paris Basin. Crosses between the European flint line and early US dent lines are adapted to northern conditions.

From each F_1 hybrid, F_5 lines were produced by classical single seed descent (SSD). So, every F_5 line can be traced back to a different F_2 plant. The three resulting populations will be denoted by X ($a \times b$), Y ($a \times c$) and Z ($b \times c$). The SSD program resulted in 129 F_5 lines for population X, 145 for population Y and 152 for population Z. One plant from each F_5 line was selfed to produce an F_6 progeny.

Evaluation of quantitative traits

In 1992, seeds from each F_6 progeny were sown in three environments which represent different latitudes between the Paris area and the south-west of France. They will be denoted by *N*(orth), *C*(entral) and *S*(outh), respectively. On each location, seeds of different F_6 progenies were planted in different rows according to a randomized design. In locations C and S rows consisted of thirty plants, in location N ten plants per row were used. In 1993, seeds from bulked $F_{7/5}$ lines were sown in location N, where each F_5 line was replicated twice using 10 seeds per replication.

The distance between plants within rows was 0.25 m and the distance between rows was 0.80 m.

In 1992 earliness was recorded as the day at which half of the plants exhibited silks. To get a more precise estimate of the delay between anther exertion and silking, data were recorded for each plant separately in the 1993 experiment. However, only silking time will be considered in this study.

Marker analysis and mapping

The lines were analyzed using RFLPs. Maps are based on:

- 98 loci of anonymous probes of the maize core map (Gardiner *et al.* 1993);
- 27 loci of Expressed Sequence Tags (sequenced cDNA for which no homology was found in gene banks; kindly provided by C. Baysdorfer, California State University);
- 60 loci corresponding to known function genes.

Analyses were carried out according to classical procedures on bulked samples of F_6 individuals.

Individual maps, as well as a joint map for all three populations, were determined using Mapmaker/EXP V3.0 (Lincoln *et al.* 1992). The synthetic map has been described by Causse *et al.* (1994, MNL 68: 42-44). Fifty-eight probes were polymorphic in the three populations.

Homogeneity of recombination frequencies across populations was investigated following Beavis *et al.* (1991). Significant differences in recombination frequencies were detected on four chromosomal segments. Most important differences in recombination frequencies observed on chromosome 8, segment UMC89-UMC30, with segment lengths of 20.2, 22.0 and 10.8 cM for populations X, Y and Z, respectively. These distances correspond to recombination frequencies of 0.34, 0.32 and 0.19, respectively.

QTL mapping per population

For each population separately data were first analyzed according to the analysis of variance (ANOVA) model:

$$Y_{ij} = \mu + G_i + R_{ij} , \qquad (1)$$

where Y_{ij} denotes the value of individual j with genotype i at the locus considered, μ denotes the grand mean of the population, G_i denotes the effect of marker genotype i and R_{ij} denotes a residual. Given the low rate of remaining heterozygosity, plants with heterozygous marker genotypes were excluded from the analysis. Application of ANOVA is restricted to marker loci only.

Subsequently, interval mapping was used to provide more precise information about the positions of QTLs. The method used is discussed by Rebaï *et al.* (1994). The method is based on model (1), where genotype effects are now written as linear combinations of regressors representing the probabilities of the genotypes at the locus considered, that are derived from flanking markers information. The method was adapted to the special case of RILs by considering the appropriate relationship between map distance and recombination frequency. Tests were performed every 2 cM.

For each position the proportion of the variation of the trait accounted for by a locus was estimated by the coefficient of determination, $R^2 = SS(\text{locus}) / SS(\text{total})$, where SS(locus) is the sum of squares associated with the locus.

In the case of ANOVA, this proportion is biased (it actually overestimates the effect of the locus), but it is possible to account for this bias (Charcosset & Gallais, submitted). However, in the case of regression interval mapping, the proportion will be biased in the opposite direction for positions at a distance from the flanking markers. Thus, specific corrections, accounting for systematic bias, should be developed.

Epistatic effects between QTLs were investigated using ANOVA. The interaction was tested for all possible pairs of loci. The percentage of significant tests (α = 0.05) was recorded for:

- all pairs of markers;
- pairs of markers for which at least one marker showed a significant effect;
- pairs for which both markers showed significant effects.

For these three situations, the percentage of significant tests was very close to what is expected under the overall null hypothesis.

Simultaneous analysis of the three populations

A global analysis was performed for the three populations using the model:

$$Y_{pij} = \mu + P_p + G_i + R_{pij} \, , \qquad (2)$$

where Y_{pij} is the value of individual j with genotype i in population p, μ denotes the grand mean, P_p denotes the effect of population p (= 1, 2, 3), G_i denotes the effect of genotype i (*aa*, *bb* or *cc*, depending on the population) and R_{pij} denotes a residual.

Model (2) can be completed by incorporating the interaction PG_{pi} between population p and genotype i:

$$Y_{pij} = \mu + P_p + G_i + PG_{pi} + R_{pij} \, . \qquad (3)$$

The interaction effect is based on one degree of freedom. The interaction effect allows testing for consistency of genotype effects across populations. To evaluate this consistency, it is possible to define the contrast:

$$E = (G_{aa}^{X} - G_{bb}^{X}) + (G_{bb}^{Z} - G_{cc}^{Z}) + (G_{cc}^{Y} - G_{aa}^{Y}) \, , \tag{4}$$

where $(G_{aa}^{X} - G_{bb}^{X})$ is the difference between the effects of genotypes *aa* and *bb* in population *X*. Assuming that no epistasis is present and that there are no differences between the recombination frequencies of the populations, the contrast E will be null. In the case of three populations, the test for the presence of interaction allows testing of the hypothesis $E = 0$.

When using ANOVA, model (3) must be restricted to the marker loci mapped in all three populations. The regression approach of Rebaï (1994) allows the use of model (3) at any position, provided that neighbouring markers supply sufficient information about the probabilities associated with the genotypes in each population. For instance, distant positions on the long arm of chromosome 9 were not considered, due to the large distance to the most informative marker in population *Z*. For model (2) effects were analysed every 2 cM. Due to the amount of computer time needed, tests for model (3) tests were only carried out at marker positions.

All computations were carried out using SAS IML programs especially developed for this study.

Results and discussion

Identification of QTLs

The effect of a given genomic region was called significant if at least one test, either in a single population or in the global analysis, was significant at the 0.5% level. This level was determined in order to get a type I error over the entire genome of approximately 10%.

Very contrasting situations were observed across the chromosomes. No QTL was detected on chromosomes 6 and 7. One QTL was detected on chromosome 4. However, it was only significant in one of the populations at only one location, and may be considered as a false positive. One QTL was detected on chromosome 3 with a moderate effect in population *Z*. On chromosomes 1, 5, 9 and 10 the *F* statistic showed two rather distant peaks (45 cM or more apart), which suggest the presence of two QTLs. Chromosomes 2 and 8, the *F* statistic showed complicated patterns, which suggest the presence of two or more linked QTLs in addition to a distant QTL. The minimum number of QTLs detected on these chromosomes is three. So, at least 15 QTLs (or distinct chromosomal segments) were detected using previously described methods and criteria.

The effect of the environment on QTL expression

Taken as a whole, results were highly consistent across locations. For instance, in population *Y*, four out of 11 QTLs showed a significant effect in the four environments, five in three environments, one in two environments, and only one QTL was specific to a single environment.

This is consistent with the high correlations which were observed across locations, which illustrates that for this range of latitudes earliness shows little genotype × environment interaction.

Comparison across populations

The number of detected QTLs appeared smaller in population *X* (7) than in populations *Y* (11) and *Z* (12). This difference becomes more important when only QTLs are considered, which are significant at the 0.5% level (rather than the 5% level): 2 in *X*, 6 in *Y* and 8 in *Z*. The numbers of detected QTLs are consistent with the differences in flowering time between the parents of the original crosses.

Using the linear model it was found that the QTL with the most important effect across all the environments was detected in population *Y* near probe UMC67 (chromosome 1) with a substitution effect of 3.6 days (15% of the variation) at location *N* in 1992. The distribution of the test statistics along chromosome 1 is presented in Figure 1 for populations *X*, *Y* and *Z*. It shows that no QTL is detected at this position in population *X*, and that in population *Z* a QTL with a much smaller effect (only significant at the 5% level) is detected at this position. This illustrates that differences in the expression of the QTLs can be observed across the populations.

The effect of genetic background on QTL expression

Tests of the effect of the genetic background on QTL expression were carried out for 14 of the 15 QTLs previously identified, of which 8 showed significant effects in at least one environment. Across the 56 (14 × 4) QTL × environment combinations, 12 (*i.e.*, 21%) showed significant effects at the 5% level.

These effects can be due to

- epistatic effects between a given QTL and the other QTLs;
- variation in recombination frequencies across the populations.

Analysis of these differences in the three RIL populations revealed that four genomic

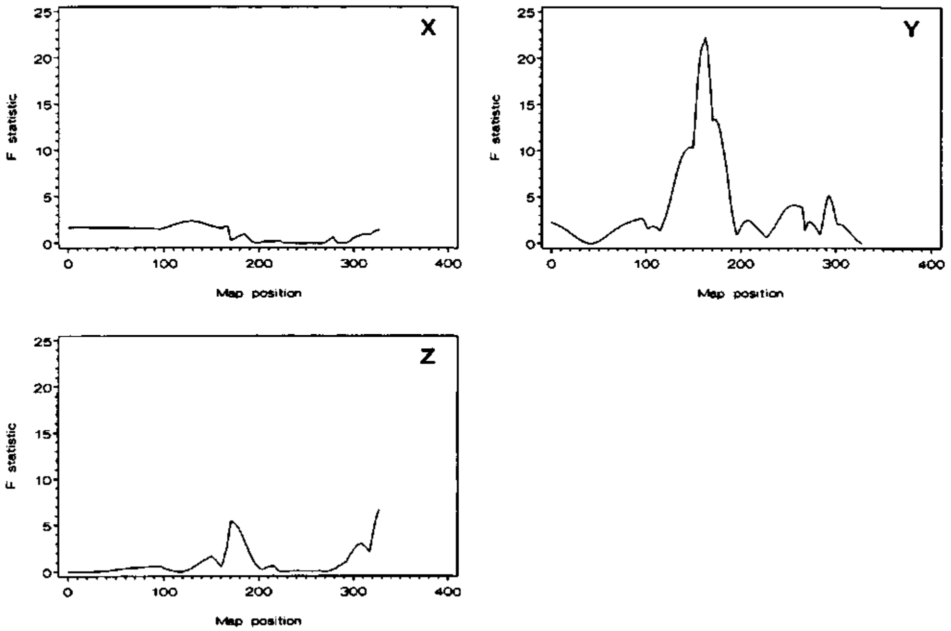


Figure 1. Analysis of the data within each population: distribution of the F statistics along chromosome 1 for populations X, Y and Z

locations displayed significant differences in recombination rates. One of these locations corresponded to a significant locus × population interaction effect.

We showed that the effect of genetic background on QTL expression can be quantified by contrast (4). If no epistatic effect are present, this contrast depends on the effects of the genotypes at the QTL(s) and on the recombination frequencies between the locus for which the test is performed and the neighbouring QTL(s). If g_{aa}, g_{bb} and g_{cc} denote the effects of the genotypes at the QTL, and R^X, R^Y and R^Z denote the recombination frequencies in populations X, Y and Z, respectively, then:

$$E = 2g_{aa}(R^X - R^Y) + 2g_{bb}(R^Z - R^X) + 2g_{cc}(R^Y - R^Z) \ . \qquad (5)$$

This expression illustrates that:

- the contrast will be very small if the distances between the QTL and the tested loci are small;

- the contrast will be very small if for distances between the QTL and the tested loci are large (because R tends to 0.5 in each population).

Most important effects are expected for intermediate distances. In this study, most important differences in recombination fractions were observed on chromosome 8, segment UMC89-UMC30, with distances of 20.2, 22.0 and 10.8 cM for populations *X*, *Y* and *Z*, respectively. These distances correspond to recombination frequencies of 0.34, 0.32 and 0.19, respectively. If these variations were observed between a marker and a QTL, they would generate a contrast equal to $2g_{aa}(0.02) + 2g_{bb}(-0.13) + 2g_{cc}(0.15)$. Considering the most important differences observed between the effects of two genotypes, *i.e.*, about 4 days, this contrast would approximately be equal to 1.12, which is much smaller than the threshold for the 5% level (approximately 2). Thus, variations in recombination rates should not be considered as a major cause for the significance of the test.

In some cases, the contrast was important when compared to genotype effects (*e.g.*, 3.5 days on chromosome 1 in environment *C* in 1992.

Conclusion

These results illustrate that several QTLs involved in variation in maize earliness show epistatic effects. No consistent epistatic effects could be detected when using the classical two-way ANOVA. However, significant effects appeared when using the simultaneous analysis presented in this paper. A possible explanation for this result is that epistatic effects between two loci are not large enough to be detected, but that the sum of the epistatic effects between one locus and all other loci may become large, and so they can be detected.

Besides of biological interest, epistatic effects should be considered when taking decisions with regard to marker-assisted breeding. The case of the QTL on chromosome 1 illustrates for instance that the transfer of the *b* allele to line *c* (population *Z*) should generate a smaller modification than expected from the results of populations *X* and *Y*. Considering the costs of marker-assisted breeding, strategies should be developed to evaluate, *a priori*, the effect of allele transfer.

Acknowledgements

We are grateful to A. Rebaï for numerous discussions and his help for the programs developed for this study. We are grateful to H. Jansen for helpful corrections on the manuscript.

References

Beavis, W.D. & D. Grant, 1991. A linkage map based on information from four F_2 populations of maize (*Zea mays* L.). Theor. Appl. Genet. 82: 636-644.

Beavis, W.D., D. Grant, M. Albertsen & R. Fincher, 1991. Quantitative trait loci for plant height in four maize populations and their associations with qualitative genetic loci. Theor. Appl. Genet. 83: 141-145.

Gardiner, J.M., E.H. Coe, S. Melia-Hancock, D.A. Hoisington & S. Chao, 1993. Development of a core RFLP map in maize using an immortalized F_2 population. Genetics 134: 917-930.

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

Lincoln S., M. Daly & E. Lander, 1992. Constructing genetic maps with Mapmaker/exp 3.0. Whitehead Institute Technical Report, 3rd Ed.

Rebaï A., B. Goffinet, B. Mangin & D. Perret, 1994. Detecting QTLs with diallel schemes. This proceedings.

Genetic diversity for allozymes, RFLPs and RAPDs in resynthesized rape

G.M. Engqvist & H.C. Becker, Department Plant Breeding Research, The Swedish University of Agricultural Sciences, S-268 31 Svalöv, Sweden

Key words

resynthesized *Brassica napus*, molecular markers, coefficients of similarity, cluster analysis, principal coordinate analysis, correlations

Summary

One way to broaden the genetic base of current oilseed rape material is to establish a gene pool consisting exclusively of resynthesized rape. We examined the genetic similarity using allozyme, RFLP and RAPD marker loci in 17 resynthesized *Brassica napus* lines with diverse parental background. Genetic similarity was estimated by the Dice coefficient for allozyme and RFLP data and by the Jaccard coefficient for RAPD data. The cluster analysis showed that the majority of lines are classified in separate groups with similarity coefficients from 0.29 to 0.85. Comparing the three molecular marker systems gave highly significant rank correlations between 0.76 and 0.53.

Introduction

The absence of wild *Brassica napus* and major emphasis on breeding for quality in the past has contributed to a rather narrow genetic base in current oilseed rape material. Therefore, there is a need of widening the genetic variation. Frequently it has been suggested to accomplish that by resynthesis of new *Brassica napus. Brassica napus* is an amphidiploid between the two diploid parents *Brassica oleracea*, kale, and *Brassica campestris*, turnip rape.

In the past resynthesized rape was mainly used to transfer one or a few favourable genes into oilseed rape by backcrossing these lines with breeding material (Engqvist & Becker 1994). A different approach to use resynthesized rape would be to improve it by recurrent selection with the outcome of a wide gene pool. Establishing a new gene pool, consisting exclusively of resynthesized rape, has several beneficial applications. First, it may broaden the genetic base of oilseed rape breeding. Second, long-time selection experiments for interesting traits can be initialized. And finally, genetically very diverse material for an efficient hybrid breeding programme can be developed.

In a comparison between oilseed rape cultivars and resynthesized lines, using allozymes and RFLPs, the latter turned out to be the one with the largest diversity (Becker *et al.* unpublished results). When investigating 23 oilseed rape cultivars by random polymorphic DNA markers Mailer *et al.* (1994) observed genetic distances below 0.2. Compared to other crops these values are very low.

Material and Methods

The 17 resynthesized *Brassica napus* lines with different parental background, presented in Table 1, have been chosen for the present investigation.

Allozyme analysis

Each accession, represented by at least two plants, was assayed electrophoretically for variation at the following seven polymorphic isozyme loci: aconitate hydratase (ACO, EC 4.2.1.3), diaphorase (DIA, EC 1.6.4.3), glucosephosphate isomerase (GPI, EC 5.3.1.9), leucine aminopeptidase (LAP, EC 3.4.11.1), 6-phosphogluconate dehydrogenase (PGD, EC 1.1.1.44), phosphoglucomutase (PGM, EC 2.7.5.1) and shikimate

Table 1. The 17 resynthesized *Brassica napus* lines used in the study

Entry	Origin	Mother	Father	Type
S3	Berlin	*B. cam. rapifera*	*B. ole. sabellica*	winter
S5	Berlin	*B. ole. capitata* 4x	*B. cam. oleifera* 4x	winter
S7	Berlin	*B. cam. oleifera* 4x	*B. ole. sabellica* 2x	winter
S9	Berlin	*B. cam. pekinensis* 2x	*B. ole. gemmifera* 2x	winter
S12	Berlin	*B. ole. sabauda* 4x	*B. cam. oleifera* 4x	winter
S13	Berlin	*B. cam. oleifera* 4x	*B. ole. medullosa* 4x	winter
S20	SvalöfAB	*B. ole. sabellica*	*B. cam. oleifera*	spring
S23	SvalöfAB	*B. oleracea*	*B. cam. oleifera*	spring
S27	DanskPlant förädling	*B. oleracea*	*B. cam. ole. annua* Nokanova	winter
S29	Göttingen	*B. ole. sabellica*	*B. cam. pekinensis*	winter
S30	Göttingen	*B. ole. capitata*	*B. cam. pekinensis*	winter
S31	Göttingen	*B. ole. italica*	*B. cam. pekinensis*	winter
S32	Göttingen	*B. ole. sabauda*	*B. cam. pekinensis*	winter
S33	Göttingen	*B. ole. alboglabra*	*B. cam. pekinensis*	spring
S40	Göttingen	*B. ole. gongylodes*	*B. cam. oleifera*	spring
S65	Berlin	*B. cam. trilocularis*	*B. ole. fructicosa*	spring
S76	Ultuna	*B. cam. oleifera*	*B. ole. botrytis*	spring

dehydrogenase (SDH, EC 1.1.1.25). DIA, GPI and SDH were analysed on starch gels as described by Becker *et al.* (1992); the remaining four enzymes were analysed on cellulose acetate according to Herbert & Beaton (1989). These isozyme systems could be encoded by 35 bands.

RFLP analysis

The RFLP analysis was carried out by Linkage Genetics, Salt Lake City, USA. A bulk-leaf sample of 20 plants was analysed for 50 nuclear DNA probes with one enzyme. In total 355 bands have been generated.

RAPD analysis

DNA from 4 plants was isolated from leaf tissue according to Edwards *et al.* (1991). The banding patterns from a selection of six primers, purchased from Operon Technologies (Alameda, Ca., USA) were obtained by performing the PCR reactions and the following sample electrophoresis as outlined in Lannér-Herrera *et al.* (1994). Bands which appeared in three replications of a RAPD run per line and primer were considered part of a RAPD profile. A total of 62 bands, out of them 18 monomorphic, of between 400 and 2000 base pairs were produced.

Statistical analysis

The banding patterns of all three marker systems were coded binary. Genetic similarity matrices based on two slightly different coefficients (Dice, Jaccard) have been calculated. For that and the following UPGMA cluster analysis and principal coordinate analysis the software package NTSYS-pc, version 1.80 (Rohlf 1993) has been used.

Measuring genetic similarity

When comparing two lines i and j, two different measures of genetic similarity GS are frequently used, the coefficient of Jaccard:

$$GS_J = \frac{N_{ij}}{(N_{ij} + N_i + N_j)} ,$$

and the coefficient of Dice, which is equivalent to 1 minus the genetic distance according to Nei & Li (1979):

$$GS_D = \frac{N_{ij}}{(N_{ij} + \frac{1}{2}N_i + \frac{1}{2}N_j)} .$$

N_{ij} is the number of bands that are common in both lines, N_i is the number of bands present in line i but not in j, and N_j is the number of bands present in line j but not in i.

As pointed out by Link *et al.* (1994), GS_J should be used for RAPD data but GS_D for RFLP and allozyme data. The arguments for this recommendation, when comparing homozygous lines, are as follows (Link, pers. comm.). RAPD marker loci have been found to mostly produce polymorphisms by either showing or not showing a band at a given gel position. Hence, one gel position represents one RAPD marker locus; a RAPD band represents one RAPD marker allele, the other alleles are represented by absent bands at this gel position. In contrast to this, allozyme and RFLP marker alleles have mostly been found to be represented by a band, but for different alleles at different gel positions. Comparing two lines, two RFLP bands together represent one RFLP locus. Therefore, if two lines differ for two RFLP bands, this reveals a difference in only one locus with two different alleles; however if two lines differ for two RAPD bands, this reveals a difference in two loci. As can be seen from the formulas given above, the use of GS_D gives only half the weight to such bands compared to the use of GS_J. The use of GS_J for RFLP and allozyme data will underestimate the genetic similarity, while the use of GS_D for RAPD data will overestimate the genetic similarity.

Nevertheless, when using the appropriate coefficient, the rank correlation between GS_J and GS_D is 1. A linear relationship between genetic similarity and genetic properties, such as the coefficient of coancestry or the degree of heterozygosity of cross-progenies, can be expected (Link, pers. comm.).

Results and discussion

A graphical presentation of the relatedness among the 17 resynthesized lines, based on the matrix generated of RAPD data using the similarity coefficient of Jaccard is given in Figure 1. The similarity coefficients range from 0.3 to 0.85 and the majority of lines are classified in separate groups. However, the cluster analysis also indicates that the relationships among some of the resynthesized lines is closer for those supplied by one institute than for those from different places. Figure 2 and 3 show the association among the 17 lines revealed by principal coordinate analysis. The x-axis separates the spring

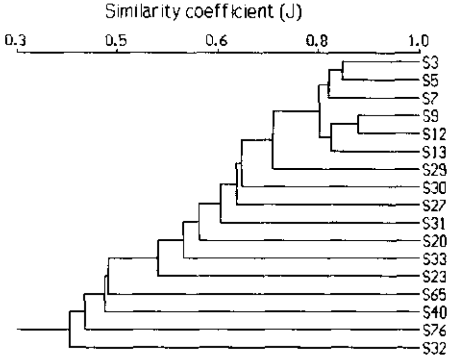


Figure 1. Dendrogram constructed from matrix of RAPD-based genetic similarity coefficients between 17 resynthesized Brassica napus lines

types from the winter types, however the RAPD-based analysis was more definitive in its separation. For both molecular markers the first two principal coordinates explain about 30% of the total variation in the data.

The rank correlations between the similarity values for all pairs of genotypes based on RAPD and RFLP were 0.76, based on RAPD and allozyme 0.67 and between RFLP and allozyme 0.53. Dos Santos *et al.* (1994) reported a similar correlation when comparing the genetic similarity between *Brassica oleracea* genotypes by RAPD and RFLP. But for example in lettuce (Landry *et al.* 1987) and maize (Messmer *et al.* 1991) it was found that these two types of molecular markers were poorly correlated. Heun et al. (1994) reported a moderate correspondence (r = 0.36) between isozyme- and RAPD-based genetic distance matrices in wild oats.

Acknowledgement

We are grateful to all colleagues who provided us with resynthesized oilseed rape

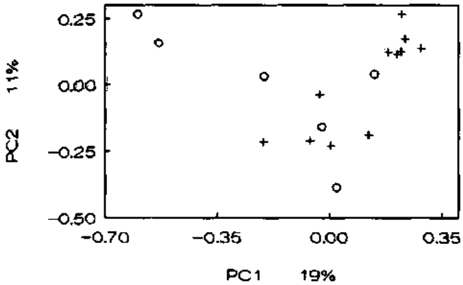
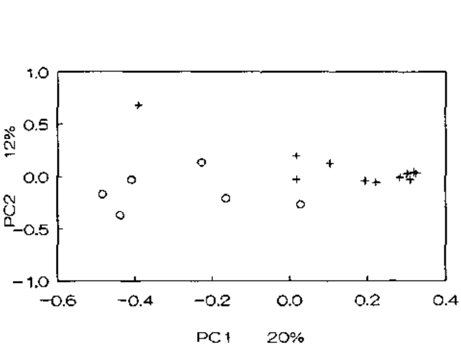


Figure 2. Relationship among 17 resynthesized *Brassica napus* lines revealed by principal coordinate analysis based on Dice similarity coefficients from RFLP data. o: spring, +: winter

Figure 3. Relationship among 17 resynthesized *Brassica napus* lines revealed by principal coordinate analysis based on Jaccard similarity coefficients from RAPD data. ○: spring, +: winter



material. Financial support from the Swedish Council for Forestry and Agricultural Research is gratefully acknowledged.

References

Becker, H.C., C. Damgaard & B. Karlsson, 1992. Environmental variation for outcrossing rate in rapeseed (*Brassica napus*) Theor. Appl. Genet. 84: 303-306.

Edwards, K., C. Johnstone & C. Thompson, 1991. A simple and rapid method for the preparation of plant genomic DNA for PCR analysis. Nuleic Acids Research 19: 1349.

Engqvist, G.M. & H.C. Becker, 1994. What can resynthesized *Brassica napus* offer to plant breeding? Sveriges Utsädförenings Tidskrift 104: 87-92.

Herbert, P.D.N. & M.J. Beaton, 1989. Methodologies for allozyme analysis using cellulose acetate electrophoresis. Helena Laboratories, Beaumont, Texas.

Heun, M., J.P. Murphy & T.D. Phillips, 1994. A comparison of RAPD and isozyme analyses for determining the genetic relationships among *Avena sterilis* L. accessions. Theor. Appl. Genet. 87: 689-696.

Landry, B.S., R. Kesseli, H. Leung & R.W. Michelmore, 1987. Comparison of restriction endonucleases and sources of probes for their efficiency in detecting restriction fragment length polymorphisms in lettuce (*Lactuca sativa* L.). Theor. Appl. Genet. 74: 646-653.

Lannér-Herrera, C., M. Gustafsson, A.-S. Fält & T. Bryngelsson, 1994. Diversity in natural populations of wild *Brassica oleracea* as estimated by isozyme and RAPD analysis. Genetic resources and crop evolution: submitted.

Link, W., C. Dixkens, M. Singh, M. Schwall & A.E. Melchinger, 1994. Genetic diversity in European and Mediterranean Faba bean germplasm revealed by RAPD markers. Theor. Appl. Genet.: in press.

Mailer, R.J., R. Scarth & B. Fristensky, 1994. Discrimination among cultivars of rapeseed (*Brassica napus* L.) using DNA polymorphisms amplified from arbitrary primers. Theor. Appl. Genet. 87: 697-704.

Messmer, M.M., A.E. Melchinger, M. Lee, W.L. Woodman, E.A. Lee & K.R. Lamkey, 1991. Genetic diversity among progenitors and elite lines from the Iowa Stiff Stalk Synthetic (BSSS) maize population: comparison of allozyme and RFLP data. Theor. Appl. Genet. 83: 97-107.

Nei, M. & M.H. Li, 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA 76: 97-107.

Rohlf, F.J., 1993. NTSYS-pc, version 1.80. Exeter Software, Setauket, New York.

Dos Santos, J.B., J. Nienhuis, R. Skroch, J. Tivang & M.K. Slocum, 1994. Comparison of RAPD and RFLP genetic markers in determining genetic similarity among *Brassica oleracea* L. genotypes. Theor. Appl. Genet. 87: 909-915.

Efficiency of marker-assisted selection

A. Gallais[1,2], *A. Charcosset*[1], [1]*Station de Génétique Végétale (INRA-UPS-INA PG), Ferme du Moulon, 91190 Gif-sur-Yvette;* [2]*Institut National Agronomique Paris-Grignon, 16 rue Claude Bernard, 75231 Paris Cedex 05, France*

Introduction

The discovery of molecular markers opens a new area for quantitative genetics and selection of complex characters. Until now quantitative genetics was the art of developing genetic laws for quantitative characters without identifying the genes involved. By tagging the genome, "invisible" genes become "visible", which makes it possible to identify, at least partially, the genotype. As a result genotypic values can be derived from the genotype at marker loci. It becomes also possible to use knowledge about the recombination between loci for accumulating favourable genes in the same genotype.

In this communication we consider the use of markers to increase the accuracy of the predictions of genotypic values. Markers cannot explain all the variation of a complex character: 40 - 70% of the total variance of an F_2 or derived population can be explained by markers for a character such as yield. It follows that the best way of predicting genotypic values is a combination of two types of information: information about the phenotype and information about the markers (Lande & Thompson 1990). Selection based on both phenotype and markers will be called marker-assisted selection (MAS). Lande and Thompson's study indicates a strong advantage of MAS over "classical" phenotypic selection, especially if the heritability is low. However, it is clear that with a low heritability, it is necessary to study more individuals. Restrictions on the number of individuals can nullify the expected advantage of MAS at low heritabilities.

The effect of such restrictions has been discussed briefly by Lande and Thompson (1990) but not with all implications for breeding methodology. By generalizing Lande and Thompson's approach, our aim is to study the effect of the number of individuals, the heritability and the proportion of genetic variance explained by markers on the relative efficiency of MAS. The theory will be developed for an arbitrary population derived from a random mating population in linkage disequilibrium. The most favourable population with high linkage disequilibrium will be that derived from an F_2. The criteria for evaluation can be of various nature: *per se* value, combining ability, S_1 value or line value.

Theory

Assumptions

To simplify the analytical approach we consider an infinite reference population. It will be assumed that the QTLs which can be marked, are independent. It is also assumed that the set of marked QTLs is independent of the set of unmarked QTLs.

Prediction of the genotypic value

With a population of infinite size, the best prediction of the genotypic value G based on phenotype P and marker information M, can be written as

$$\hat{G} = E[G \mid P,M] = G^* + b(P-G^*) \ ,$$

G^*, the genotypic value predicted by the markers, can be written as $G^* = \sum_i g_i^*$, where $g_i^* = \mathrm{E}[P \mid M_i]$ (i = 1, 2, ... , l; l is the number of marked QTLs). Lande and Thompson (1990) call g_i^* the molecular score of QTL i. It is assumed that G^* and (P-G^*) are independent and follow a normal distribution with zero mean.

It results from regression theory and the above independence assumptions that

$$b = \frac{\mathrm{cov}(G,P)-\mathrm{cov}(G,G^*)}{\mathrm{var}(P-G^*)} = \frac{\sigma^2_G-\sigma^2_{G^*}}{\sigma^2_P-\sigma^2_{G^*}} = \frac{(1-m^2)h^2}{1-m^2h^2} \ ,$$

where $\sigma^2_{G^*}$ is the genetic variance explained by the markers, $m^2 = \sigma^2_{G^*}/\sigma^2_G$ and $h^2 = \sigma^2_G/\sigma^2_P$, the heritability in the broad sense for predicting genotypic values of individuals.

The problem is that neither b nor G^* are known. So, in a selection experiment they have to be replaced by estimates $\hat{b}$ and $\hat{G}^*$. In that case the predictor becomes

$$\hat{G} = \hat{G}^* + \hat{b}(P-\hat{G}^*).$$

Assuming that G and $\hat{G}$ are bivariate normal, the expected genetic advance becomes

$$\Delta G = i \ \mathrm{cov}(G,\hat{G})/\sqrt{\mathrm{var}\ \hat{G}}\ .$$

To study the factors affecting ΔG, expressions for $\mathrm{cov}(G,\hat{G})$ and $\mathrm{var}(\hat{G})$ will be replaced by their expectation over all possible experimental populations of size N which can be drawn from the reference population. To simplify the approach it will be assumed that the coefficient b is known without error. This means that we consider the predictor for G at the level of the reference population and replace G^* by $\hat{G}^*$. The predictor thus obtained may be compared with the classical predictor used for phenotypic selection: $\hat{G} = h^2\ P$. Considering the expression for b this means that the heritability h^2 must be known with high precision from other experiments. Furthermore, it has been shown by simulation that the effect of errors in m^2, the proportion of the genetic variance explained by markers, on the relative efficiency of MAS are relatively small in comparison with the effect of errors in $\mathrm{cov}(G,\hat{G}^*)$ (Hospital & Gallais, unpublished data). Obviously, the error in m^2 has a larger effect on the accuracy of the genetic advance.

With the above assumptions it is possible to obtain

$$\mathrm{E}[\mathrm{cov}(G,\hat{G})] = (1-b)\ \mathrm{E}[\mathrm{cov}(G,\hat{G}^*)] + b\ \sigma^2_G\ .$$

It can be derived that

$$\mathrm{E}[\mathrm{cov}(G,\hat{G}^*)] = \sum_i \sigma^2_{g_i^*} + \frac{1}{\tilde{n}} \sum_i (\sigma^2_G - \sigma^2_{g_i^*}) = \sum_i \sigma^2_{p_i^*}\ ,$$

where $\tilde{n}$ is the equivalent number of plants per class of marker genotype, $\tilde{n} = (N-1)/(c-1)$, and c is the number of marker classes (c = 3 for an F_2-derived population, c = 2 for recombinant inbred or DH lines).

In the same way it can be derived that

$$\mathrm{E}[var(\hat{G})] = b^2\sigma^2_P + (1-b)^2\mathrm{E}[\mathrm{var}(\hat{G}^*)] + 2b(1-b)\mathrm{E}[\mathrm{cov}(\hat{G}^*,P)]\ ,$$

where

$$\mathrm{E}[\mathrm{var}(\hat{G}^*)] = \sum_i \sigma^2_{p_i^*} + \sum_i \sum_{j\neq i} \sigma_{p_i^* p_j^*}$$

and

$$\mathrm{E}[\mathrm{cov}(\hat{G}^*,P)] = \sum_i \sigma^2_{p_i^*}\ .$$

For full details see Charcosset & Gallais (1994).

To simplify further, we consider the case where all QTLs have the same effect:

$$\sum_i \sigma^2_{g_i^*} = l\sigma^2_{g^*}\,, \quad \sum_i \sigma^2_{p_i^*} = l\sigma^2_{p^*}\,, \quad \sum_i \sum_{i \neq j} \sigma_{p_i^* p_j^*} = l(l-1)\sigma_{p_1^* p_2^*}\,.$$

It remains to calculate $\sigma_{p_1^* p_2^*}$. This is a complex calculation which, by the assumption of the independence of the marked QTLs, tends towards zero when $\tilde{n}$ is large. A minimum value is $\frac{1}{N-1}(\sigma^2_{g_1} + \sigma^2_{g_2})$.

Prediction of offspring value

When offspring are evaluated according to the same system of test as the parents (the case of direct selection), the same derivations can be made. It is only necessary to change the meaning of G and g_i^*: G becomes the genotypic value of the offspring and g_i^* becomes the additive value of the parents at marked QTL i. Note that additive values can be estimated directly, without any specific design, by using the markers in the parent generation.

Numerical application

After simplification of previous expressions, only m^2, h^2, l and $\tilde{n}$ affect ΔG. This makes it possible to study the effect of these four parameters. To simplify the calculations we will consider a complete linkage between markers and QTLs. The biological situation considered is the following: l independent marked QTLs are present, which explain a proportion m^2 of the genetic variance. In this preliminary study, a marker will be entered into the predictor of the genotypic value if it is detected by ANOVA using a 5 % significance level. However, it will be better to consider groups of markers; the one-marker approach will detect less QTLs and, consequently, will decrease the expected efficiency of MAS.

So, we have to consider the distribution of the number of QTLs detected (with values ranging from 0 to l) with regard to repeated samples from the reference population. Obviously, this number is determined by the power of the experiment, and in general it will be smaller than the theoretical maximum l. It may even be much smaller for low heritabilities, if the percentage of variance explained by markers is small, if $\tilde{n}$ is small or if the number of QTLs is large (and consequently QTL effects are small). The determination of the power for a particular experiment is made according to results from Charcosset and Gallais (1994). The relative efficiency (RE) is computed from the ratio

$$RE = \frac{\Delta G_{\mathrm{MAS}}}{ih\sigma_{\mathrm{G}}} .$$

Results from the numerical application

The results given here are for a population of doubled haploids derived from an F_1 (c = 2). An F_2-type population (c = 3) has also been considered. The corresponding detailed results are not given, because they can be derived from the case c = 2. For the same values of the parameters, *RE* is decreased by approximately 15%. For given values of h^2, m^2 and the number of marked QTLs, *RE* remains unchanged if the total number of genotyped plants is multiplied by 1.5.

Effect of m^2 on RE for given N, l, h^2

For $h^2 > 0.40$ the increase in *RE* will generally be small or, even, negligible. For low heritabilities, it appears that *RE* increases nearly linearly with increasing m^2, and, for realistic values of *N*, *RE* is much smaller the values reported by Lande and Thompson (1990) for an infinite number of plants. For example, for $h^2 = 0.15$, $l = 10$ and $m^2 = 0.50$, *RE* is increased by only 40% ($N = 500$) instead of 90% ($N = \infty$). Limitations on the numbers of plants studied strongly decreases *RE*. For $h^2 = 0.15$, $l = 10$ and $N = 200$, *RE* = 1.12. Increasing the number of QTLs, thereby decreasing QTL effects, also decreases *RE*. It appeared that it is important to consider the ratio $l/\tilde{n}$. When this ratio is larger than 0.05, *RE* < 1.25 for $m^2 < 0.60$.

Effect of h^2 for given m^2, N and l

RE decreases rapidly when h^2 increases. *RE* is generally larger for low heritabilities. However, when for very low heritabilities the probability of detecting QTLs is small for N < 500 and $l/\tilde{n} > 0.02$, and as a consequence *RE* tends to 1 or may even become smaller than 1. This means that there is a value of h^2 for which *RE* is a maximum for given values of *N* and $l/\tilde{n}$. The optimum value of h^2 lies 0.10 and 0.20. Note that for $N = 200$, $l = 5$ ($l/\tilde{n} = 0.025$) *RE* is about 1.50, *i.e.*, there is an increase in efficiency of 50%.

Note that MAS can be worse that non-assisted selection when ghost QTLs (false-positive QTLs) are used in the prediction. The probability of such a situation increases at low heritabilities, small values of *N* and and small QTL effects. This problem has not been considered here; it will contribute to decreasing *RE*.

Effect of the number of marked QTLs and the ratio l/n

With the assumption of equal effects for marked QTLs, the ratio $l/\tilde{n}$ for given h^2 and m^2 strongly determines *RE*. It appears that generally for realistic numbers of plants, it will not be possible to use many marked QTLs in the prediction; 10 is a maximum for $N < 500$, *i.e.*, for $l/\tilde{n} < 0.02$. With $l/\tilde{n} = 0.02$, $h^2 = 0.10$ and $m^2 = 0.65$, *RE* is about 1.60. This values reduces to 1.40 if $h^2 = 0.30$. With $l/\tilde{n} = 0.05$, *RE* lies in the range from 1.15 to 1.20. It is obvious that for given m^2, *RE* is a maximum for the smallest possible number of marked QTLs, *i.e.*, 1. Consequently, for the same value of m^2, spreading the total QTL effect over a number of QTLs will decrease *RE*.

Combined effect of h^2 and m^2 for given l and N

The domain of values of h^2 and m^2 where *RE* is larger than 1.2 consists of relatively small values of h^2 and large values of m^2. If we limit m^2 to 0.65, then this range is either empty or very small for $l/\tilde{n} > 0.05$. If $l/\tilde{n} < 0.05$ (*e.g.*, $l = 5$, $N = 200$ or $l = 10$, $N = 400$) the range increases, especially if $l/\tilde{n}$ becomes smaller.

It appears clearly from the previous results that the gain in efficiency due to MAS can be very small in a large range of situations: medium to high heritabilities and medium to small proportions of genetic variance explained by the markers.

Comparison with other breeding methods

MAS versus other ways of increasing heritability

For a given selection intensity the increase in efficiency of selection by using markers is produced by an increase of the heritability. This is due to the fact that information from the markers is known with great precision. Consequently, for the breeder it is interesting to know whether the use of markers is competitive compared with other means of increasing the heritability, mainly the use of replications of the genotypes, the use of associated characters and the use of correlations among relatives.

The use of replications

It appears that for the same heritabilities 8-10 replications give about the same gain in efficiency as MAS if 50-60% of variance is explained by markers. For a complex character it will be difficult to explain more of 50% of the genetic variance by using markers, so the breeder has the choice between investing in markers or in more replications. However, considering that the space at the disposal of the breeder is limited, the increase in the number of replications will generally be at the expense of the number of studied units, which results in a decrease in selection intensity. Then MAS offers the possibility to improve efficiency without strong limitations on selection intensity.

The use of associated (secondary) characters

It appears that character-assisted-selection (CAS) can be competitive with MAS if less than 50% of the genetic variance is explained by markers. Note that MAS is a particular case of CAS; it is very similar to multitrait-assisted-selection. The main difference is that usually it will be easier to find informative markers than to find informative characters.

The use of correlations among relatives

At low heritabilities, the use of kinship through S_1 or FS combined selection can be competitive with MAS if 30 to 50% of genetic variance is explained by markers. However, to be competitive combined selection must involve relatively large family sizes (n = 20), a situation which is not favourable for a good long term management of genetic variability.

Efficiency of MAS according to breeding method

Three methods have been considered: mass selection, full-sib family selection and half-sib progeny selection. For family or progeny selection three replications were considered with 20 individuals per plot, and an environmental correlation between plants on the same plot of 0.50. It appears that for low broad sense heritabilities ($h^2 < 0.15$) with the same number of genotyped plants the three methods have about the same *RE*. However, for FS and HS selection we have to consider heritabilities at the level of family means. Taking such an heritability equal to 0.50 gives an *RE* of 1.15 to 1.20 if 50% of the genetic variance can be explained by markers. Then MAS, as expected, is not very efficient for methods with family or progeny testing except at very low heritabilities and with a low number of plants per family.

It is worthwile to emphasize that mass and FS selection are more efficient if the proportion of the genetic variance explained by markers is small. This is due that by the markers, it is possible to estimate directly the additive effect of M-QTLs. HS selection is not so affected because at the level of the phenotype it is already a selection on additive value; the effect of dominance appears only through the limited number of individuals per progeny.

A consequence of these results is that if classical mass selection is less efficient than other methods for low heritabilities ($h^2 < 0.10$), the use of markers can change this well known conclusion.

Conclusions

From this study of the relative efficiency of marker-assisted selection it appears mainly that the use of markers can be efficient in all cases where the heritability is low due to the presence of high environmental and dominance variances. MAS may provide an interesting improvement of mass selection. It may also increase the efficiency of methods using family or progeny testing through a decrease of the required number of replications and an increase of the selection intensity. It must be also noted that if all advantages of increasing heritability by MAS are nullified because it is possible to increase heritability by other means, methods involving markers may still be preferred because of a better management of genetic variability and control of recombination.

Obviously the cost of MAS must be considered in comparison to non-assisted selection. However, considering cost of MAS is not sufficient when MAS leads to a greater genetic advance: it is necessary to consider the whole breeding strategy of the firm.

Finally, it must be noted that even without complete linkage between markers and QTLs, it is possible to carry out selection using markers only without phenotypic re-evaluation. With the use of off-season generations this can contribute to an increase of the genetic advance per unit of time if cost of genotyping are low (only efficient markers must be studied). The problem will be to preserve the genetic variability on the unmarked part of the genome.

References

Charcosset, A. & A. Gallais, 1994. Estimation of the contribution of a marker to the variance of a quantitative trait. Theor. Appl. Genet.: submitted.

Lande, R. & R. Thompson, 1990. Efficiency of Marker-Assisted Selection in the improvement of quantitative traits. Genetics 124: 743-756.

Selection of markers linked to quantitative trait loci by regression techniques

Christine A. Hackett, Scottish Agricultural Statistics Service, Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA, U.K.

Introduction

Quantitative traits are determined by the joint effects of several quantitative trait loci (QTLs) and the environment. The presence of a QTL will cause the genotype classes of nearby markers to have different means for that trait and this may be represented as a regression of the trait value on the marker genotype. Similarly the effects of several QTLs on a trait can be represented by a multiple regression model. However, a linkage map consists of a large number of markers and, for a given trait, many will be uninformative. Techniques are needed to scan a set of markers rapidly and select those most closely associated with the trait. Cowen (1989) has briefly described the possibilities of different regression procedures for selecting markers and Romero-Severson *et al.* (1989) have studied the selection of markers in an F_2 cross using best subset regression. Jansen (1993) uses backward stepwise regression to select markers prior to interval mapping of multiple QTLs. Here several variable selection techniques are examined for their ability to select the correct markers and some model checking procedures are discussed.

Materials and methods

Consider a population of doubled haploid lines derived from a cross between two inbred lines which differ with respect to a quantitative trait and genetic markers. If a QTL with alleles Q_1 and Q_2 is linked, with recombination fraction θ, to a marker with alleles M_1 and M_2 the trait value T_i of line i (= 1, 2, ... , n; n is the number of doubled haploid lines) may be expressed by the equation

$$T_i = \alpha + \beta x_i + \varepsilon_i \ , \qquad (1)$$

where x_i is 0 or 1 for marker genotypes M_1M_1 or M_2M_2. Let the trait means associated with QTL genotypes Q_1Q_1 and Q_2Q_2 be μ_1 and μ_2. Then the regression coefficient, β, is equal to the difference between the means of the two marker genotype classes:

$$\beta = (1 - 2\theta)(\mu_2 - \mu_1) . \tag{2}$$

When $\theta = 0.5$, the regression coefficient $\beta = 0$. However if $\theta < 0.5$ then β differs from zero and the significance of this may be tested by the usual t-test or F-test. The effects of more than one QTL may be similarly modelled by a multiple regression equation

$$T_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_n x_{ni} + \varepsilon_i . \tag{3}$$

A linked marker N, lying between M and Q, will have a smaller recombination fraction with Q and hence a larger value of β, a larger F statistic and a larger value of R^2, the proportion of variation explained. The closest marker to the QTL among a linked group will be expected to have the highest value of R^2.

Variable selection methods

A multiple regression model such as equation (3), using all available markers, would have many non-significant coefficients. We aim to select a few markers which are closely linked to QTLs for each trait and to exclude markers which do not contribute substantially. Possible selection methods are stepwise regression (and the related methods of forward selection and backward elimination) and best subset regression. In stepwise regression a sequence of regressions is computed by adding or dropping a marker at each step, according to the significance of the regression coefficient. Forward selection starts with no markers present and adds markers with significant coefficients. Backward elimination starts with all markers present and drops those with non-significant coefficients. In best subset regression there are computer algorithms for determining the 'best K' subsets which compute only a fraction of the total number of possible regressions. Subsets may be assessed using the maximum adjusted coefficient of determination R_a^2 or Mallows' C_p. Mallows' C_p statistic is defined for a regression on p parameters by

$$C_p = \frac{SSE_p}{s^2} - (n - 2(p + 1)) , \tag{4}$$

where s^2 is the error mean square from the equation containing all the markers (Mallows 1973). The best equation should have a low value of C_p which is close to $p + 1$.

Influential observations

An ideal regression model should not depend largely on a few observations. Cook's distance (Cook 1977) is a measure of the influence of the i^{th} observation:

$$D_i = (y^P - y^P_{-i})^t\,(y^P - y^P_{-i})/Ps^2\ , \tag{5}$$

where y^P are the predicted values from a regression equation including the full set of P markers and y^P_{-i} are the predicted values when the i^{th} observation is omitted. Léger & Altman (1993) discuss the influence of observations on the variable selection procedure and develop a modified version of Cook's distance for this situation:

$$D_i^u = (y^p - y^{(p)}_{-i})^t\,(y^p - y^{(p)}_{-i})/ps^2\ , \tag{6}$$

where y^p are the predicted values from a regression on the p selected markers and $y_{-i}^{(p)}$ are from the model selected when the i^{th} observation is omitted. It is necessary to repeat the variable selection n times, removing one of the n observations in each case.

Simulation study

Twenty data sets, consisting of 100 doubled haploid lines and 18-24 loci were simulated to lie on three or four chromosomes: $\{A_1\text{-}A_5\}$, $\{B_1\text{-}B_7\}$, $\{C_1\text{-}C_6\}$ and $\{D_1\text{-}D_6\}$. Distances between markers were randomly selected from a uniform distribution between 0 and 40 cM. Three quantitative traits were simulated as the equal, additive effects of four loci:

$$\begin{aligned} X &= 10 \times (A_5 - B_4 + C_1 + C_5) - 10,\\ Y &= 10 \times (A_4 - A_2 + B_6 + C_6) - 10,\\ Z &= 10 \times (A_4 + A_2 + B_6 + C_6) - 20. \end{aligned} \tag{7}$$

Different degrees of random variation were added ($N(0,2^2)$, $N(0,5^2)$, $N(0,10^2)$ and $N(0,20^2)$) to give four traits in each set. The traits are denoted by X_1, X_2, ... , Z_4. The four quantitative trait loci were excluded from the subsequent regression analysis.

Experimental study

The set of real data consisted of 59 doubled haploid lines derived from a cross between the spring barley cultivar Blenheim and the SCRI spring barley breeding line E224/3. DNA extracted from the plants was assessed for 84 molecular markers, mainly RAPD markers but also a few RFLPs. Chalmers *et al.* (1993) used bulked segregant analysis to identify RAPD markers linked to milling energy in this cross. The data is reanalysed here to demonstrate the regression method.

Results

Analysis of simulated data

Table 1 summarises the results of the different selection procedures on the three sets of traits. The number of QTLs for which a flanking marker is selected decreases with the heritability of the trait. It is also lower for traits in set Y, where the linked QTLs of opposite effects are more frequently undetected. The proportion of regressions where the linked QTLs are not detected also increases as the heritability decreases. Most linked markers are selected using the maximum R_a^2 criterion, but this method also selects the most spurious markers. The backward elimination method selects more spurious markers than forward or stepwise regression: this is generally due to models derived from the backward regression including pairs of neighbouring markers, with opposite signs, to improve the fit for the few observations at which the markers have different genotypes. Stepwise and forward selection generally give the same model. All the markers selected by stepwise regression and backward elimination have significant coefficients but some of the markers selected by the C_p or maximum R_a^2 criteria had non-significant coefficients.

We now give two examples in more detail, to look for influential observations. In one set of simulated data, stepwise, forward and backward selection and best subset

Table 1. Comparison of the different selection methods, averaged over the 20 simulations. S = stepwise regression, F = forwards selection, B = backwards selection, C_p = best subsets regression with C_p criterion, R_a^2 = maximum adjusted R^2

Trait	Mean no. flanking markers correctly selected					Mean no. linked pairs of QTLs not detected					Mean no. spurious markers selected				
	S	F	B	C_p	R_a^2	S	F	B	C_p	R_a^2	S	F	B	C_p	R_a^2
X_1	3.8	3.8	3.9	3.8	3.9	0.0	0.0	0.0	0.0	0.0	0.3	0.5	0.9	0.6	2.7
X_2	3.6	3.6	3.8	3.8	3.8	0.1	0.1	0.0	0.0	0.1	0.5	0.6	0.9	0.6	2.4
X_3	2.8	2.7	3.1	3.2	3.6	0.4	0.5	0.5	0.4	0.2	0.2	0.3	0.6	0.3	1.6
X_4	1.3	1.3	1.5	1.7	2.7	0.9	0.9	0.8	0.8	0.4	0.2	0.2	0.5	0.7	2.1
Y_1	3.4	3.4	3.5	3.4	3.7	0.2	0.2	0.1	0.2	0.0	0.4	0.4	0.6	0.4	1.8
Y_2	3.0	3.0	3.5	3.5	3.6	0.5	0.5	0.3	0.3	0.1	0.3	0.4	0.9	0.3	1.7
Y_3	2.0	2.0	2.2	2.3	3.5	0.7	0.7	0.6	0.6	0.2	0.2	0.3	0.4	0.3	1.7
Y_4	0.9	0.9	1.0	0.9	2.6	1.0	1.0	1.0	1.0	0.5	0.1	0.1	0.3	0.2	1.5
Z_1	3.7	3.7	3.8	3.6	3.9	0.1	0.1	0.1	0.1	0.0	0.3	0.4	0.9	0.8	2.3
Z_2	3.6	3.6	3.6	3.6	3.9	0.2	0.2	0.2	0.1	0.0	0.3	0.3	1.0	1.0	2.3
Z_3	2.8	2.8	2.8	2.8	3.5	0.4	0.4	0.4	0.5	0.1	0.2	0.2	0.5	0.2	1.6
Z_4	1.8	1.8	1.8	2.0	3.2	0.9	0.9	0.9	0.9	0.4	0.3	0.3	0.4	0.3	2.0

regression for Z_4 using the C_p criterion all selected the same subset of variables, A_3, C_5, D_5, C_1 and C_2. All of these had coefficients with $p < 0.05$. The maximum Cook's distance was 0.1, indicating no observations had a large influence on the fit of this model. The model with maximum R_a^2 also contained markers C_3 and B_4, but the coefficients of these were not significant and their inclusion reduced the significance of C_5. Marker C_5 flanks the QTL on chromosome C and marker A_3 lies between the two QTLs on chromosome A, but markers D_5, C_1, C_2 and C_3 are spurious selections. However, this would not be known in a set of real data. C_1 and C_2 are linked with a recombination fraction $r = 0.13$ and have regression coefficients of similar sizes but opposite signs.

The stepwise regression was repeated 100 times, removing each observation in turn. 63 runs selected the same subset of observations as the full set, $\{A_3, C_5, D_5, C_1, C_2\}$. 5 runs selected the set $\{A_3, C_5, D_5, C_2\}$ and 31 runs selected the set $\{A_3, C_5, D_5\}$. The remaining run selected the set $\{A_1, C_5, A_3\}$. The pair of markers C_1 and C_2 are sensitive to the set of observations used, indicating that this is probably a spurious linkage. Markers A_3 and C_5 are selected in every case. The observation which led to the set $\{A_1, C_5, A_3\}$ had the largest modified Cook's distance of 3.1. This has an extreme value among the observations with crossovers between A_1 and A_3, but does not appear as an outlier among the full set of observations. The dependence of D_5 on the presence of this observation suggests this marker may not be linked to a QTL.

In the second example, stepwise, forward and backward selection for X_4 indicated two markers, B_6 and D_4, which is not actually linked to a QTL. No observation had a high Cook's distance in this model. However, when the highest observation was dropped the model changed to B_5 and C_6 (both of which are next to QTLs). The modified Cook's distance for this observation was 7.5. Again, a single observation has a large influence on the variable selection procedure.

Analysis of the milling energy data

Stepwise and forward selection gave the same five markers: OPD13-H900, OPE11-H400, OPB10-H, OPB4-H300 and OPA19-H2000. Backward elimination replaced OPE11-H400 by a closely linked marker and best subset regression (best C_p) omitted OPA19-H2000. The model with the maximum R_a^2 included two extra variables, both with non-significant regression coefficients and one with the opposite sign to its sign in a single regression equation. Hence these two markers probably are spurious. The influence of individual observations was investigated by dropping each observation in turn and repeating the stepwise regression. The first four markers were consistently included but OPA19-H2000 was frequently omitted. Marker OPD13-H900, which accounted for the largest individual percentage of the variation, is in the group linked to Rrn2 which is known to lie on chromosome 5H (Chalmers *et al.* 1993). In the regression equation markers OPD13-H900 and OPE11-H400 have negative signs i.e. the Blenheim alleles at the QTLs linked to these markers are associated with high milling energy while those linked to markers OPB10-H and OPB4-H300 are associated with low milling energy. The importance of the marker OPB4-H300 was not revealed by an analysis using MAPMAKER, as no other markers linked to it were available.

Discussion

Variable selection methods have been shown to be a useful tool in the preliminary screening of a set of markers to identify those closely linked to QTLs for a trait, especially when linked QTLs are involved. The markers are selected in an approximate order of importance, while neighbouring markers giving no additional information are excluded. Examination of the trends along the chromosome of the percentage variation accounted for by individual regressions will indicate the approximate position of a single QTL but is unlikely to detect two linked QTLs. If more than one order of the markers is possible different locations might be inferred for the QTL. The multiple regression analysis is unrelated to the ordering of the markers and may be used to explore a data set early in a mapping project before a marker map is available.

Variable selection methods are available in many statistical packages but do need to be used with caution. One problem is when to stop to exclude unlinked markers. Subsets selected by the best C_p criterion sometimes contained markers whose regression coefficients were not significantly different from zero at the 5% level. Selected markers whose individual regression coefficients are not significant, or are of the opposite sign to their coefficients in individual regressions, should also be examined carefully. The influence of individual observations on the variable selection procedure should be investigated; this is time-consuming but may give valuable information.

These techniques have been applied here to data on doubled haploid plants. However, they can equally be applied to backcross data or adapted to the case of F_2 offspring by using two variables to represent the additive and dominance effect of each marker (Edwards *et al.* 1987). We have also used these methods for a preliminary screening of RAPD markers in a population of dihaploid potato lines from two heterozygous parents.

The regression coefficients are functions of the recombination fractions and the QTL genotype means and hence these two effects cannot be estimated separately. However once closely linked markers have been identified other types of model, such as normal mixture models, may be used to estimate the parameters of these effects (*e.g.*, Weller 1986, Knapp *et al.* 1990). Knapp *et al.* (1990) have also used regression models on indicator variables representing the four genotype classes defined by two flanking markers, with coefficients specified as explicit nonlinear functions of the recombination fractions and QTL genotype means. However, Knapp (1991) stresses that where multiple QTL affect a trait, all parameters should be estimated simultaneously. The use of regression models to describe the location of a QTL relative to two flanking markers has also been examined by Haley & Knott (1992) and Martinez & Curnow (1992), who have both examined the biases which can result if the possibility of linked QTLs is ignored.

Acknowledgements

I thank W.T.B. Thomas for the milling energy scores, the Cellular and Molecular Genetics Department for the use of their molecular data from the Blenheim E224/3 barley cross, and J.W. McNicol and W. Powell for their suggestions on the manuscript. The work was supported by funds from the Scottish Office Agriculture and Fisheries Department.

References

Chalmers, K.J., U.M. Barua, C.A. Hackett, W.T.B. Thomas, R. Waugh & W. Powell, 1993. Identification of RAPD markers linked to genetic factors controlling the milling energy requirement of barley. Theor Appl Genet 87: 314-320.

Cook, R.D., 1977. Detection of influential observation in linear regression. Technometrics 19: 15-18.

Cowen, N.M., 1989. Multiple linear regression analysis of RFLP data sets used in mapping QTLs. In T. Helentjaris & B. Burr (Eds.) Development and application of molecular markers to problems in plant genetics. Current Communications in Molecular Biology, pp. 113-116. Cold Spring Harbour Laboratory Press.

Edwards, M.D., C.W. Stuber & J.F. Wendel, 1987. Molecular-marker-facilitated investigations of quantitative-trait loci in maize. I. Numbers, genomic distribution and types of gene action. Genetics 116: 113-125.

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

Jansen, R.C., 1993. Interval mapping of multiple quantitative trait loci. Genetics 135: 205-211.

Knapp, S.J., W.C. Bridges Jr. & D. Birkes, 1990. Mapping quantitative trait loci using molecular marker linkage maps. Theor Appl Genet 79: 583-592.

Knapp, S.J., 1991. Using molecular markers to map multiple quantitative trait loci: models for backcross, recombinant inbred and doubled haploid progeny. Theor Appl Genet 81: 333-338.

Léger, C. & N. Altman, 1993. Assessing influence in variable selection problems. Journal of the American Statistical Association 88: 547-556.

Mallows, C.J., 1973. Some comments on C_p. Technometrics 15: 661-675.

Martinez, O. & R.N. Curnow, 1992. Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. Theor Appl Genet 85: 480-488.

Romero-Severson, J., J. Lotzer, C. Brown & M. Murray, 1989. Use of RFLPs for analysis of quantitative trait loci in maize. In T. Helentjaris & B. Burr (Eds.) Development and application of molecular markers to problems in plant genetics. Current Communications in Molecular Biology, pp. 97-102. Cold Spring Harbour Laboratory Press.

Weller, J.I., 1986. Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers. Biometrics 42: 627-640.

Effects of heterozygosity and heterogeneity on the adaptation of sorghum [*Sorghum bicolor* (L.) Moench] to a semi-arid area of Kenya

Bettina I.G. Haussmann & Hartwig H. Geiger, University of Hohenheim, Institute of Plant Breeding, Seed Science, and Population Genetics (350), D-70593 Stuttgart, Germany

Key words

drought, heterosis, individual/populational buffering, sorghum

Abstract

Twenty four parent lines, 12 single-cross hybrids, 12 two-component blends of parent lines, and 12 two-component hybrid blends of grain sorghum [*Sorghum bicolor* (L.) Moench] were grown in eight macro-environments (site/season combinations) in a semi-arid area of Kenya. Environmental means for grain yield ranged from 47 to 570 $\mathrm{g\,m^{-2}}$, reflecting a wide range of drought patterns and stress intensities. In all environments, hybrids significantly outyielded their homozygous parent lines, with a mean relative heterosis of 53%. Blending effects were small and inconsistent. Averaged over all environments and both levels of heterozygosity, the blending effect was zero. A logarithmic transformation of the original data was undertaken to reduce non-additivity among genetic and environmental effects. The transformation did not result into homogeneous error variances. Combined across environments, genetic differences were significant only among lines. Genotype by environment interaction effects were much more important than genotypic effects. Genotypes reacted differently to preflowering and terminal drought stress. Lines in pure stand displayed a greater genotype by environment interaction than the other three groups. Stability (Eberhart & Russell 1966) was not associated with mean grain yields, and entries with below average reaction to drought stress were found within all four types of genetic structure. In conclusion, hybrids displayed a much higher yielding potential and a slightly improved phenotypic stability compared to their parent lines.

Introduction

Sorghum belongs to the major crops of the semi-arid tropics of Africa and Asia with mean grain yields of 80 and 120 $g\,m^{-2}$, respectively (FAO Production Yearbook 1992). Semi-arid areas of Kenya are characterised by low and erratic rainfall. Although terminal drought stress is prevailing, dry spells can occur anytime during the growing season, leading to an unpredictable drought stress pattern. Local farmers usually do not have access to irrigation facilities and totally rely on the adaptability and yield stability of their rainfed crop varieties.

Adaptedness to unpredictable drought stress may include drought escape, dehydration avoidance, and dehydration tolerance (Blum 1988). It can occur on the level of single genotypes and on the level of genetically heterogeneous plant populations. Allard & Bradshaw (1964) coined the terms individual and populational buffering to distinguish between the two phenomena when reviewing possible mechanisms of yielding stability in variable environments. Individual buffering may be favoured by heterozygosity, and populational buffering by heterogeneity in as much as the different genotypes present in the population are specifically adapted to different environmental conditions (Allard & Bradshaw 1964, Bradshaw 1965, Reich & Atkins 1970, Schnell & Becker 1986, Becker 1987, and others).

Reich & Atkins (1970) studied the effects of population type (lines, line blends, hybrids, hybrid blends) on grain yield of sorghum in Iowa, United States of America (USA). In their study, environmental means for grain yield ranged from 400 to 740 $g\,m^{-2}$. Mean relative heterosis amounted to 25%. Other studies conducted in the USA gave similar estimates, *e.g.*, Kambal & Webster (1966) and Patanothai & Atkins (1974), 20 and 22% mean relative heterosis, respectively. Contrasting to these investigations, Jowett (1972) found a much higher (88%) superiority of hybrids over lines and varieties under growing conditions of East Africa with an environmental mean of 228 $g\,m^{-2}$. Mean blending effects were estimated at 2% in the study of Reich & Atkins (1970). Ross (1965) reported a 1% superiority of hybrid blends over the mean of their pure stands. Accordingly, results from an experiment grown in northern Ghana indicated no yield advantage of two- to five-component mixtures of local cultivars over the individual stands (Mercer-Quarshie 1979). Yet, Bebawi & Abdelaziz (1983) reported a 25% superiority of two-component blends of varieties differing in their maturity dates over the mean of their pure stands in a study conducted in Sudan under furrow irrigation with environmental means of 141 to 417 $g\,m^{-2}$.

Regarding stability of grain yield, Reich & Atkins (1970) reported hybrid blends to be the most productive and stable population type although none of the populations was distinctly superior for all parameters. Jowett (1972) as well as Patanothai & Atkins (1974) compared three-way crosses with single crosses. Considerable variation among individual hybrids for stability parameters was found, suggesting that stability of performance may be attainable with either single or three-way crosses. Francis *et al.* (1984) found hybrids to be more stable than open-pollinating varieties in early planting while the reverse was true in late planting. Mercer-Quarshie (1979) reported a trend of increasing stability with increasing complexity of the mixtures.

The objective of this study was to investigate the effects of heterozygosity and heterogeneity on yield and yield stability of sorghum in an extremely variable semi-arid area of Kenya.

Materials and methods

The tested genotypes consisted of twelve unrelated single cross hybrids and their respective 24 parent lines (Table 1) and represent actual breeding materials from SADC/ICRISAT (Southern African Development Community/International Crop Research Institute for the Semi-Arid Tropics) Zimbabwe and ICRISAT India. The following four types of genetic structure were formed, analogous to the experiments of Reich & Atkins (1970) and Schnell & Becker (1986):

1. Homogeneous entries of homozygous plants (24 parent lines, using maintainers in case of cytoplasmic-genic male sterile lines);
2. Homogeneous entries of heterozygous plants (12 hybrids);
3. Heterogeneous entries of homozygous plants (12 two-component blends of parent lines according to the parentage of the hybrids);
4. Heterogeneous entries of heterozygous plants (12 two-component blends of hybrids such, that each hybrid was represented in two mixtures.

Disregarding maternal effects the arrangement is genetically balanced in that all four types of population have the same content of nuclear genes. Entries were divided into

Table 1. Designation and pedigrees of the hybrids used in this study

Hybrid	Pedigree			Hybrid	Pedigree		
SDSH-409	Ma-6	×	R-8602	SDSH-300	ICSA-20	×	SDS-170
SDSH-19	ATx-623	×	SDS-3219	SDSH-48	ICSA-12	×	SDS-6013
ICSH-110	ICSA-296	×	ICSR-33	SDSH-339	ATx-631	×	A-6352
SDSH-315	ICSA-21	×	R-8609	SDSH-4	D2-A	×	SDS-3880
SDSH-215	SPL-23A	×	MR-855	SDSH-343	A-150	×	SDS-2690
ICSH-205	ICSA-51	×	ICSR-152	SDSH-398	A-8607	×	ZAM-1518

two sets. The two sets were each planted together with six check varieties in 6×6 triple lattice designs. Plots consisted of three to four rows, three to four meters long, resulting in a plot size of 9.6 to 12.8 m^2. The spacing between rows was 0.8 m and between plants within rows 0.2 m. To ensure a true 1:1 mixture in the heterogeneous entries, the whole experiment was hill planted by hand and the components of each blend were sown alternately in the successive hills of each row. Thinning was done to one plant per hill aiming at a final plant density of 6.25 plants m^{-2}.

The experiment was grown in eight macro-environments (site/season combinations) in the Makueni District, Kenya, during 1991 to 1993 (Table 2). The total amount of water received by the single experiments ranged from 151 to 1078 mm, including supplemental irrigation for stand establishment given in some environments. Several traits were assessed but only data of grain yield [$g\,m^{-2}$, 9.5 to 10% grain moisture] is being considered here. Grain yields were linearly corrected in the event of bird or squirrel damage; no adjustment for missing plants was undertaken.

The computer program PLABSTAT (Utz 1991) was used for statistical analyses. In a first step, data of each set were analyzed according to the lattice design with extreme outliers (Anscombe & Tukey 1963) declared as missing values. Phenotypic correlations among environments were calculated by using lattice-adjusted mean values of each entry from the individual environments. Homogeneity of error variances was tested with Bartlett's Test (Snedecor & Cochran 1980). Combined analyses of variance across environments including stability analyses (Eberhart & Russell 1966) were computed with logarithmically transformed data (Transformation: Y'= ln [(grain yield in $g\,m^{-2}/10)+1$] since genetic and environmental effects were related in a multiplicative manner, indicated by Tukey's test for non-additivity (1949). In general, non-additivity is apparent when materials of very different yield potential are compared (Jowett 1972, Becker 1987).

Table 2. Site/season combinations and amount of water [mm] received by the single experiments; SR = short rainy season, LR = long rainy season

Location	Season			
	SR 1991-92	LR 1992	SR 1992-93	LR 1993
Kibwezi A	430	259	1078	173
Kibwezi B	373			
Kibwezi C		275		
Kiboko			597	151

A = Irrigation Project, B = Goat Research Station, C = Local Farm

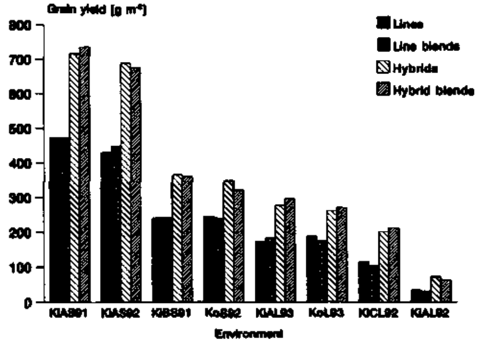


Figure 1. Means of the four types of genetic structures in the eight environments for grain yield, with Ki and Ko referring to the locations Kibwezi and Kiboko, A, B, and C to the three sites within Kibwezi, and S91, S92, L92, and L93 to the rainy seasons SR 1991-92, SR 1992-93, LR 1992, and LR 1993, respectively

Results and discussion

Environmental means for grain yield ranged from 48 to 584 g m^{-2} (Figure 1), reflecting the wide range of drought patterns and stress intensities. In all environments, hybrids significantly outyielded their parent lines. Mean relative heterosis was 53%, and an increase of relative heterosis was observed in the two most severely stressed experiments. The estimated mean relative heterosis lies within the range given in the literature. Compared to data obtained in the USA, estimates from studies conducted in Africa are much higher. This might be explained by the fact that female parents used in African hybrid breeding programs usually derive from US lines and thus are not well adapted to Africa (Majisu & Doggett 1972).

Blending effects were small and inconsistent with both lines and hybrids, and on the overall average, the blending effect was zero. Possible reasons for lack of blending effects in our study are: materials did not differ extremely in developmental or morphological traits. Apart from limited stemborer [*Chilo partellus* (Swinhoe), *Sesamia calamistis* Hmps.], shootfly [*Atherigona soccata* (Rondani)] and charcoal rot [*Macrophomina phaseolina* (Tassi) Goid] infestation in single environments, no marked diseases or pests were encountered. Therefore, the potential advantages of genetic heterogeneity, such as decrease of intergenotypic competition and reduced spread of pests and diseases could not materialize.

As a first indicator of genotype by environment interaction, phenotypic correlations among environments differing in the kind and degree of stress were calculated for grain yield (Table 3). The two environments characterized by preflowering and respectively moderate terminal drought stress (but with similar environmental means) were tightly correlated with the non-stress environment but only moderately with each other, indicating interaction between entries and the kind of stress. Only weak or non-

Table 3. Phenotypic correlations among environments differing in drought stress for grain yield

Drought stress	No stress	Preflowering	Moderate terminal
Preflowering	0.79 **		
Moderate terminal	0.74 **	0.52 **	
Severe terminal	0.39 **	0.20 ns	0.39 **

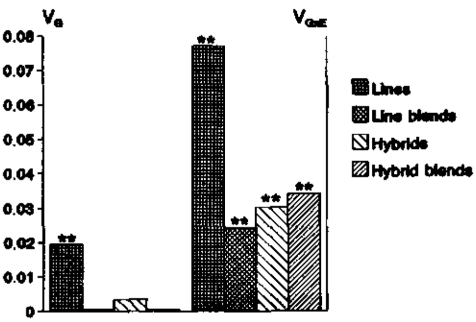
** Significant at the 0.01 probability level; ns non-significant

significant relationships existed between the extreme stress environment and the other environments. From this we may conclude that under most extreme conditions only specialists can survive and that these specialists may not be the highest yielding genotypes in the other environments. The effect of the experimental error on the correlations should be small since mean values averaged over three replications were correlated. However, an upward bias could have resulted from the consistent superiority of hybrids over lines.

The logarithmic transformation sharply reduced non-additivity, but error variances still remained heterogeneous so that the F-Tests in the combined analyses were only approximate (Cochran & Cox 1957). The estimates of genetic variance were significant only among lines (Figure 2). Genotype by environment interaction variances were much more important than genetic variances in all four groups. Genotype by environment interaction was highest among lines in pure stand. Relative to lines in pure stand both heterozygosity and heterogeneity led to a reduction of genotype by environment interactions. Heterogeneity, however, did not reduce genotype by environment interaction at the heterozygous level. Effective populational buffering may require more diverse hybrids or more complex mixtures than those evaluated in the present study.

Heterogeneity of regressions explained 33 and 41% of the genotype by environment interaction sums of squares in the two sets of materials, respectively. The rather low fit of the linear model may be explained by the fact that environments with similar

Figure 2. Estimates of the genetic (V_G) and the genotype by environment interaction component of variance ($V_{G\times E}$) in the four types of genetic structure, logarithmic scale (** Significant at the 0.01 probability level)



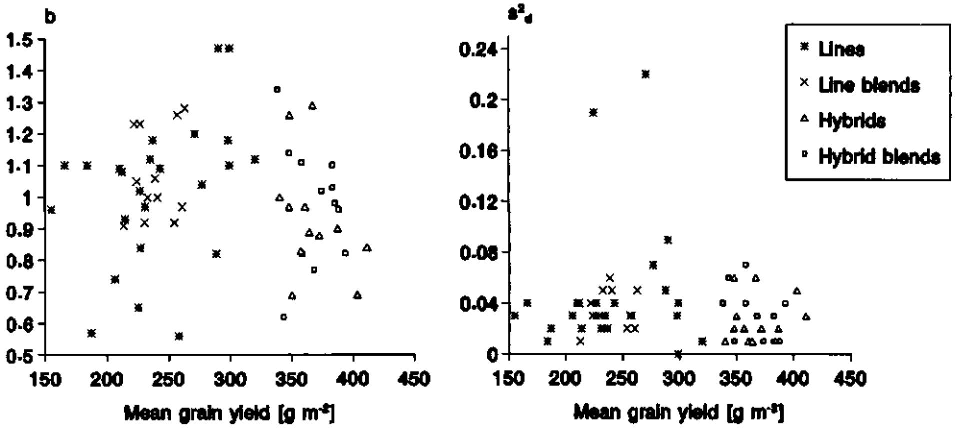


Figure 3. Relationship between mean and stability parameters for grain yield (logarithmic scale; left: regression coefficient (b); right: deviation mean square (s_d^2))

environmental means differed in the kind of stress, *i.e.*, that each environment represented a unique combination of several stress factors, and that genotypes reacted differently to these changing environmental conditions. The extreme environmental variation resulting from different drought patterns and stress intensities could not be adequately expressed in the one-dimensional environmental index.

Regression coefficients were not associated with mean grain yields (Figure 3). Large values of the regression coefficient, calculated on a logarithmic scale, indicate a very sharp proportional decline of yield under severe stress, a situation that plant breeders should be anxious to avoid in stress prone environments (Jowett 1972). Overall, the ranges of the regression coefficient were wide among lines, hybrids and hybrid blends, but reduced among line blends. On average, heterozygotes had a slightly lower regression coefficient.

Regarding deviation mean squares (Figure 3), two lines had outstandingly high values, *i.e.*, strongly varied around the average response. Principally, the deviation mean squares showed the same trend as the regression coefficients: ranges within population types were high whereas differences between means of the four groups were tiny with hybrids in pure stand having the lowest value. It should therefore be possible to select stable entries (small deviation from regression) with below average reduction of grain yield under increasing stress conditions (regression coefficients below one) at both levels of heterozygosity and heterogeneity.

In summary, heterozygosity turned out to be an important prerequisite for obtaining high grain yields of sorghum grown in semi-arid climates, particularly under severe drought stress. The extreme genotype by environment interactions indicate that heterogeneous cultivars should display a higher potential for yielding stability than uniform entries. However, no direct evidence was found for this supposition.

Acknowledgements

The financial support granted by GIARA (German Israel Agricultural Research Agreement for Developing Countries) is highly appreciated. Technical advice, store capacity, seeds and/or land within their farms or Research Stations were kindly provided by ICRISAT, the University of Nairobi, Kenya Agricultural Research Institute, Kibwezi Irrigation Project and Anthony Mutie. Furthermore, we greatfully acknowledge the excellent technical assistance of Daniel, Ziro, and many others.

References

Allard, R.W. & A.D. Bradshaw, 1964. Implications of genotype-environmental interactions in applied plant breeding. Crop Sci. 4: 503-507.

Anscombe, F.J. & J.W. Tukey, 1963. The examination and analysis of residuals. Technometrics 5: 141-160.

Bebawi, F.Z. & A.H. Abdelaziz, 1983. Grain sorghum responses to pure stands and mixtures under irrigation. Trop. Agric. (Trinidad) 60: 262-264.

Becker, H.C., 1987. Möglichkeiten zur züchterischen Verbesserung der Ertragssicherheit bei Getreide. Habilitationsschrift, Fakultät III - Agrarwissenschaften I der Universität Hohenheim, Stuttgart, April 1987.

Blum, A., 1988. Plant Breeding for Stress Environments. CRC Press Inc., Boca Raton, Florida.

Bradshaw, A.D., 1965. Evolutionary significance of phenotypic plasticity in plants. Adv. in Genetics 13: 115-155.

Cochran, W.G. & G.M. Cox, 1957. Experimental Designs. Second Edition. John Wiley & Sons, Inc., London, New York.

Eberhart, S.A. & W.A. Russell, 1966. Stability parameters for comparing varieties. Crop Sci. 6: 36-40.

Francis, C.A., M. Saeed, L.A. Nelson & R. Moomaw, 1984. Yield stability of sorghum hybrids and random-mating populations in early and late planting dates. Crop Sci.: 1109-1112.

Jowett, D., 1972. Yield stability parameters for sorghum in East Africa. Crop. Sci. 12: 314-317.

Kambal, A.E. & O.J. Webster, 1966. Manifestations of hybrid vigor in grain sorghum and the relations among the components of yield, weight per bushel, and height. Crop Sci. 6: 513-515.

Majisu, B.N. & H. Doggett, 1972. The yield stability of sorghum varieties and hybrids in East African environments. East Afr. Agric. Forest. J., October 1972: 179-192.

Mercer-Quarshie, H., 1979. Yield of local sorghum (*Sorghum vulgare*) cultivars and their mixtures in northern Ghana. Trop. Agric. (Trinidad) 56: 125-133.

Patanothai, A. & R.E. Atkins, 1974. Yield stability of single crosses and three way hybrids of grain sorghum. Crop Sci. 14: 287-290.

Reich, V.H. & R.E. Atkins, 1970. Yield stability of four population types of grain sorghum, *Sorghum bicolor* (L.) Moench, in different environments. Crop Sci. 10: 511-517.

Ross, W.M., 1965. Yield of grain sorghum (*Sorghum vulgare* Pers.) hybrids alone and in blends. Crop Sci. 5: 593-594.

Schnell, F.W. & H.C. Becker, 1986. Yield and yield stability in a balanced system of widely differing population structures in *Zea mays* L. Plant Breeding 97: 30-38.

Snedecor, G.W. & W.G. Cochran, 1980. Statistical methods. Seventh edition. The Iowa State University Press, Ames, IOWA, USA.

Tukey, J.W., 1949. One degree of freedom for non-additivity. Biometrics 5: 232-242.

Utz, H.F., 1991. "PLABSTAT". A computer program for the statistical analysis of plant breeding experiments. Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, Germany.

Mapping of quantitative trait loci by using genetic markers: an overview of biometrical models used

Ritsert C. Jansen, Centre for Plant Breeding and Reproduction Research (CPRO-DLO), P.O. Box 16, 6700 AA Wageningen, The Netherlands

Introduction

In crop plants quantitative variation is a feature of many important traits, such as yield, quality or disease resistance. Means of analyzing quantitative variation and especially of uncovering its potential genetic basis are therefore of prime importance for breeding purposes. It has been demonstrated in the early 20th century that such quantitative variation results from the combined action of multiple segregating genes and environmental factors (Johannsen 1909). An intrinsic feature of such traits is, however, that the individual genes contributing to quantitative variation can hardly be distinguished. The genetics of such complex traits is therefore studied in general terms (population means and variances, covariances between progenies, heritabilities and so on) of classical quantitative genetics (Mather & Jinks 1971), rather than in terms of individual gene effects. Only by the use of genetically marked chromosomes, is it possible to detect and locate the loci affecting quantitative traits ("quantitative trait loci" or "QTLs"). Linkage between QTLs and morphological markers (Sax 1923; Rasmusson 1933; Thoday 1961) has been reported, but accurate and systematic genetic mapping has been hampered by the lack of a sufficient number of genetic markers covering an entire genome. Recently, new tools have become available by the advent of molecular markers, such as restriction fragment length polymorphisms (RFLPs) (Botstein *et al.* 1980, Beckmann & Soller 1983). Now, dense genetic linkage maps exist for many plant and animal species, which heralds a new era for quantitative genetics (Tanksley *et al.* 1989).

Powerful and accurate biometrical methods are needed, so as to make possible the dissection of quantitative variation of complex characters into individual QTL effects. Mapped QTLs can be traced in breeding programmes, for instance, indirectly by selection for linked markers, or they can be cloned and introgressed via molecular or cell-biological techniques. The traditional methods for mapping of QTLs are, however, neither powerful nor accurate and the development of better methods is an area open to research. Not surprisingly, the detection and mapping of QTLs is gaining rapidly growing attention from biometrical geneticists.

Biometrical models

Here, we give a short overview of the advancements in biometrical modelling of the QTL mapping problem. The models will be briefly described for backcross progenies, but the same ideas also apply to other types of progeny, in which linkage association between markers and QTLs is manifest.

Studying single markers one by one

The traditional approach to detecting and mapping QTLs involves studying single markers one by one (Sax 1923, Soller & Brody 1976). Allele substitution effects at a marker locus indicate the presence of one or more linked QTLs. In the case of a backcross progeny, the expected difference between the two marker classes, say Mm and mm, is:

$$\mu_{Mm} - \mu_{mm} = \sum a_i(1-2r_i) \ , \tag{1}$$

where the summation is over QTLs, r_i is the recombination frequency between the marker and the i^{th} QTL, and a_i is the allele substitution effect of the i^{th} QTL. The realized value of $1-2r_i$ is likely to be close to 0 for unlinked QTLs (unless the progeny size is small), and the effect of those QTLs is negligible. The F-test in analysis of variance is commonly used to test for the allele substitution effect at the marker locus. It is assumed that $Y=\mu_{Mm}+E$ for individuals in marker class Mm, and $Y=\mu_{mm}+E$ for individuals in marker class mm, where Y is the value of the phenotypic trait and E is a random normally distributed error. In short regression notation:

$$Y = \mu_{mm} + x(\mu_{Mm}-\mu_{mm}) + E \ , \tag{2}$$

where the indicator variable x takes the value 0 and 1 for the genotypes mm and Mm, respectively, and $\mu_{Mm}-\mu_{mm}$ is the allele substitution effect.

This marker-one-by-one approach has a number of shortcomings. In the case of a single segregating QTL, (a) tight linkage to a single QTL with a small effect cannot be distinguished from loose linkage to a single QTL with a large effect; (b) the position of a single QTL relative to the marker is not defined accurately. In the case of multiple QTLs, (c) the method is not powerful since QTLs are mapped one a time, ignoring the effects of other mapped QTLs; (d) the method cannot separate linked QTLs; (e) effects of QTLs with opposite sign effects cancel so that the test for the allele substitution effect at a marker locus is not even a proper test for QTL activity; (f) the presence of QTLs with effects of equal sign can lead to the false detection of a single "ghost-QTL" at an intermediate marker; Finally, (g) the error distribution is actually a mixture of (normal) distributions (due to recombinations between the marker and QTLs; see below).

Mixture models for a single QTL with one or two flanking markers

Weller (1986) emphasized that the trait should be considered to follow a mixture of (normal) distributions and he developed mixture models for estimating the linkage between a single marker and a single QTL. Suppose that F_1 individuals with genotype MQ/mq are backcrossed to the parent with genotype mq/mq. For individuals in marker class Mm the model is $Y=\mu_{Qq}+E$ when no recombination between the marker and the QTL has occurred (chance $1-r$), and $Y=\mu_{qq}+E$ otherwise (chance r). Similarly, for individuals in marker class mm, the model is $Y=\mu_{qq}+E$ when no recombination between the marker and the QTL has occurred (chance $1-r$) and $Y=\mu_{Qq}+E$ otherwise (chance r). In short regression notation:

$$Y = \mu_{qq} + X(\mu_{Qq}-\mu_{qq}) + E\ , \tag{3}$$

where $\mu_{Qq}-\mu_{qq}$ is the allele substitution effect at the QTL and X is a random indicator variable which takes values 0 and 1 for the genotypes qq and Qq, respectively, with probabilities r or $1-r$ depending on the marker genotype. If the phenotypic values are not affected by a QTL, then $Y=\mu+E$, *i.e.*, $\mu_{Qq}=\mu_{qq}=\mu$. The test for the presence of a putative QTL is commonly based on a comparison of the likelihood of the model with the QTL and that of the model without the QTL (the likelihood-ratio test).

Weller's approach has been generalized so as to make possible the analysis of single QTLs enclosed by a pair of flanking markers (Simpson 1989, Lander & Botstein 1989, Jensen 1989, Knapp *et al.* 1990). This flanking marker procedure has been termed "interval mapping". The regression model (3) is still used, but the distribution of X now depends on the two flanking markers. Expressions for the (conditional) probabilities of the various genotypes can be derived straightforwardly.

The interval mapping method has several advantages over the traditional approach. In the case of a single segregating QTL, (a) the location and the effect of the QTL can be assessed more accurately; (b) the likelihood for the presence of a putative QTL can be plotted along the genetic map, so as to present the evidence for QTLs at the various positions of the genome; (c) the test for the presence of a QTL is more powerful. The principal shortcoming of interval mapping is that still only models for a single QTL are used, which is in clear contradiction with the commonly assumed oligogenic or polygenic nature of quantitative traits. Therefore, interval mapping has a number of shortcomings when two or more QTLs are segregating; see the points (c)−(f) listed in the previous section. This has motivated theoretical research for multiple QTL mapping methods.

Standard multiple regression of the trait on the markers

The simple method based on regression of phenotype on markers one by one has been generalized to multiple regression methods in which the trait can be regressed on a large number of markers (Cowen 1989, Stam 1991, Rodolphe & Lefort 1993, Jansen 1993, Zeng 1993, Jansen & Stam 1994). If the marker map sufficiently covers the whole genome, the major part of the QTL induced variation will be absorbed by marker cofactors. The regression model reads:

$$Y = \mu + \sum x_i a_i + E , \tag{4}$$

where the summation is over marker loci, and x_i and a_i are the indicator variable and the allele substitution effect for the i^{th} marker, respectively. Individuals with any missing marker observation might be eliminated from the regression, but in regression of the trait on many markers only a very limited set of data would then remain. Jansen & Stam (1994) developed the exact model, *i.e.*, a mixture model, in which the indicator variable x_i is replaced by a random indicator variable X_i, the probability distribution of which is based on the observations at the linked marker loci (see below). Rodolphe & Lefort (1993) replaced the indicator variable x_i by the expectation of X_i given the observations at linked marker loci.

The multiple regression approach has several clear advantages: (a) the background "noise" is reduced (but not minimized) by taking into account the effects of QTLs by nearby markers; (b) by starting with a 'polygenic' model (regression on all markers) it gets around detection and mapping problems with interfering QTLs; (c) in regression on all markers, the test for QTL activity in a certain region is generally unaffected by QTLs that are located in other regions; (d) standard procedures for selection of important variables in regression can be used, so as to identify the "important" markers, hopefully those flanking the QTLs. Compared to interval mapping, the multiple regression approach has the disadvantage that (a) no precise information for the QTL location or the QTL effect is obtained and (b) no QTL likelihood plots are produced. Further, (c) in regression on all markers, the test for QTL activity is not powerful due to genetic correlation between the QTL and markers outside the region under study; (d) the overall significance level in QTL detection is unclear when standard selection methods are used.

Multiple regression models based on the expected values of the marker class means

Several authors (Knapp *et al.* 1990, Knapp 1991, Haley & Knott 1992, Martinez & Curnow 1992, Moreno-Gonzalez 1992) have developed similar approximate interval mapping methods, which could be generalized so as to map several QTLs simultaneously. These models are based on the expected phenotypic values of the marker classes, which are non-linear functions of QTL effects and recombination frequencies. The interval mapping model given by expression (3) is approximated by the model:

$$Y = \mu_{qq} + \mathscr{E}_M(X)(\mu_{Qq} - \mu_{qq}) + E \,, \tag{5}$$

i.e., X in expression (3) is replaced by its expectation $\mathscr{E}_M(X)$, given the observed genotype at the flanking marker loci. For multiple QTLs the regression model reads:

$$Y = \mu + \sum \mathscr{E}_M(X_i)a_i + E \,, \tag{6}$$

where the summation is over putative QTLs; the variables X_i are the indicator variables for the QTLs, and the a_i are the allele substitution effects of the QTLs. Knapp *et al.* (1990) and Knapp (1991) ignore double and multiple crossovers to simplify the model. They estimate the recombination parameters in the non-linear models by direct means. Like in the interval mapping method, Haley & Knott (1992) and Martinez & Curnow (1992) move the QTL along the chromosome, and at each map location the likelihood for the presence of a putative QTL is plotted. At a given map location the recombination frequencies are known (and with that $\mathscr{E}_M(X)$), so that expression (5) is a standard regression model with unknown parameters μ_{Qq} and μ_{qq}. This approach can be generalized to a two-dimensional search for two QTLs (by moving independently two QTLs along the chromosomes) or to a multidimensional search for multiple QTLs (by moving independently multiple QTLs along the chromosomes). To simplify the models, Moreno-Gonzalez (1992) ignores double crossovers between flanking markers and locates putative QTLs at a fixed position, namely halfway between their flanking markers. This makes it possible to regress the trait on many QTLs in a way similar to standard multiple regression of the trait on markers (in which case putative QTLs are "located at marker positions"). The models of Moreno-Gonzalez are, however, much more complex.

The advantages of these methods compared to interval mapping are: (a) the effects of linked QTLs can be unravelled more efficiently and more accurately; (b) when two QTLs are simultaneously searched for, the simultaneous likelihood for the presence of these QTLs can still be plotted in a three-dimensional graph; (c) the computer programme is easy and fast. There are, however, several disadvantages: (a) the complexity of the models increases with the number of putative QTLs in the model; (b) the computation involved with all these models is almost unfeasible when the number of QTLs is larger than two or three; (c) two or three putative QTLs can be moved simultaneously along the chromosomes but other (mapped or not yet mapped) QTLs will be ignored; (d) the random variable X for the QTL in the mixture model is replaced by its expected value, but this approximation is not efficient in the case of major QTLs or QTLs located in the middle of wide marker intervals.

Mixture models and approximate mixture models for multiple QTLs

Jansen (1992) developed exact models for multiple QTLs. We number the loci (markers and putative QTLs) according to their map order; X_i is the indicator variable for the i^{th} locus. The regression model reads:

$$Y = \mu + \sum X_i a_i + E \ , \tag{7}$$

where the summation is over putative QTLs. Jansen (1992) demonstrated how the simultaneous likelihood of the trait (Y), the QTLs (X_i) and their flanking markers (X_{i-1} and X_{i+1}) can be maximized; in fact it was demonstrated that the mixture model can easily be embedded in the framework of multiple linear regression models and even in that of generalized linear models. The problem can be considered as a multiple regression problem with missing genetic data. The core of the method is to augment and complete the data: in case of a single QTL all data are replicated twice; the first replication is completed with the QTL genotype qq, the other replication with Qq, and corresponding weights (conditional probabilities) can be calculated. Parameter estimation is carried out by iterative weighted regression of the augmented data on the QTLs, alternating updating of the weights and updating of the parameter estimates. If many QTLs are assumed, the number of possible genotypes becomes so large that computation is no longer feasible. Disregarding genotypes with negligible weights can be a solution, without substantial loss of information.

Jansen (1992) described a "hybrid" method, combining interval mapping with standard multiple regression methods (see also Jansen (1993) and Zeng (1994)). The regression model reads:

$$Y = \mu_{qq} + X(\mu_{Qq} - \mu_{qq}) + \sum X_i a_i + E \; , \tag{8}$$

where X is the random indicator variable for the single QTL, and the summation is over markers used as cofactors. Jansen & Stam (1994) developed a very general method of multiple linear regression of a quantitative trait on genotype (QTLs and markers). This regression model is the same as that in expression (7), but now the summation is over loci in general, *i.e.*, over QTLs and over those markers used as cofactors. Here, the method will be termed "MQM mapping", where MQM is an acronym for "multiple-QTL models" as well as for "marker-QTL-marker", which reflects the insertion of QTLs between markers on the genetic map. The basic idea is the completion of any missing genotypic (QTL or marker) data by augmenting and weighting the data. Marker observations can be fortuitously missing, but also other types of missing marker data occur in a natural way. For instance in an F_2, when markers are dominant and the heterozygote cannot be distinguished from one of the homozygotes. Or in outbred progeny, when markers with different information are located in mixed order on the chromosomes (only one of the gametes gives information on recombination if a marker segregates according to backcross rules, whereas both gametes are informative if a marker segregates according to F_2 rules). Jansen (1994) studied the chance of type I or type II errors in MQM mapping.

Advantages of the models for MQM mapping are: (a) the full power of complete linkage maps is exploited as much as it is computationally feasible, to complete any missing genetic (QTL and marker) data; (b) the likelihood for the presence of a putative QTL can be plotted along the genome when marker cofactors are used; (c) Models, which are exact for major QTLs and approximate for minor QTLs, can be fitted.

Concluding remarks

We have sketched the recent developments of QTL mapping methods from the traditional marker-one-by-one approach, via the "single QTL" interval mapping approach to more advanced methods based on exact or approximate models for multiple QTLs. Presently the traditional marker-one-by-one approach and the interval mapping method are still widely used (*cf.* Paterson *et al.* 1991, Stuber *et al.* 1992, De Vicente & Tanksley 1993). But it is now generally recognized that simultaneous mapping of multiple QTLs is more efficient and more accurate. Therefore, the methods based on simultaneous mapping of multiple QTLs should provide the method of choice for the analysis of QTL mapping data. These methods date, however, from the past two years and their properties are still being studied analytically or by simulation.

References

Beckmann, J.S. & M. Soller, 1983. Restriction fragment length polymorphisms in genetic improvement methodologies, mapping and costs. Theor. Appl. Genet. 67: 35-43.

Botstein, D., R.L. White, M. Skolnick & R.W. Davis, 1980. Construction of a genetic map in man using restriction length polymorphisms. Am. J. Hum. Genet. 32: 314-331.

Cowen, N.M., 1989. Multiple linear regression analysis of RFLP data sets used in mapping QTLs. *In:* Helentjaris T, B. Burr (Eds.) Development and application of molecular markers to problems in plant genetics. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, pp. 113-116.

De Vicente, M.C. & S.D. Tanksley, 1993. QTL analysis of transgressive segregation in an interspecific tomato cross. Genetics 134: 585-596.

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

Jansen, R.C., 1992. A general mixture model for mapping quantitative trait loci by using molecular markers. Theor. Appl. Genet. 85: 252-260.

Jansen, R.C., 1993. Interval mapping of multiple quantitative trait loci. Genetics 135: 205-211.

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

Jansen, R.C., 1994. Controlling the type I and type II errors in mapping quantitative trait loci. Genetics: in press.

Jensen, J., 1989. Estimation of recombination parameters between a quantitative trait locus (QTL) and two marker gene loci. Theor. Appl. Genet. 78: 613-618.

Johannsen, W., 1909. Elemente der exakten Erblichkeitslehre. Fisher, Jena.

Knapp, S.J., W.C. Bridges & D. Birkes, 1990. Mapping quantitative trait loci using molecular marker linkage maps. Theor. Appl. Genet. 79: 583-592.

Knapp, S.J., 1991. Using molecular markers to map multiple quantitative trait loci: models for backcross, recombinant inbred, and doubled haploid progeny. Theor. Appl. Genet. 81: 333-338.

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

Martinez, O. & R.N. Curnow, 1992. Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. Theor. Appl. Genet. 85: 480-488.

Moreno-Gonzalez, J., 1992. Genetic models to estimate additive and non-additive effects of marker-associated QTL using multiple regression techniques. Theor. Appl. Genet. 85: 435-444.

Paterson, A.H., S. Damon, J.D. Hewitt, D. Zamir, H.D. Rabinowitch, S.E. Lincoln, E.S. Lander & S.D. Tanksley, 1990. Mendelian factors underlying quantitative traits in tomato: comparison across species, generations, and environments. Genetics 127: 181-197.

Rasmusson, J.M., 1933. A contribution to the theory of quantitative character inheritance. Heriditas 18: 245-261.

Rodolphe, F. & M. Lefort, 1993. A multi-marker model for detecting chromosomal segments displaying QTL activity. Genetics 134: 1277-1288.

Sax, K., 1923. Association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. Genetics 8: 552-560.

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

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

Stam, P., 1991. Some aspects of QTL analysis. In: Proceedings of the eighth meeting of the Eucarpia section

"Biometrics in plant breeding", BRNO.

Stuber, C.W. S.E. Lincoln, D.W. Wolff, T. Helentjaris & E.S. Lander, 1992. Identification of genetic factors contributing to heterosis in a hybrid from two elite maize inbred lines using molecular markers. Genetics 132: 823-839.

Tanksley, S.D., N.D. Young, A.H. Paterson & M.W. Bonierbale, 1989. RFLP mapping in plant breeding: new tools for an old science. Biotechnology 7: 257-264.

Thoday, J.M., 1961. Location of polygenes. Nature 191: 368-370.

Weller, J.I., 1986. Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers. Biometrics 42: 627-640.

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

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

Towards an automated interpretation of gel electrophoresis

Zivan Karaman, Limagrain Genetics Biometrics Unit, B.P. 115, 63203 Riom Cedex, France, Phone: (33) 73 63 43 43, FAX: (33) 73 63 43 39,
E-mail: karaman@cicc.univ-bpclermont.fr

Key words
density function, genetic fingerprinting, RFLPs, software

Abstract
The RFLP data should theoretically be in the form of a small number of discrete bands called variants (which are often considered as alleles), but since the observed data are subject to different sources of noise and form a set of continuous values, a scoring procedure must be used to assign a discrete value to each band. When a huge quantity of data must be analysed, hand scoring is error-prone and difficult, if not impossible. We sought an automated scoring procedure to avoid errors that could be introduced by hand scoring. The problem was addressed using the non-parametric probability density estimate of the distribution of molecular weights for each probe, and looking for its local maxima. This gives us a set of possible discrete values; each observed data point is assigned to the nearest discrete value by calculating all the distances and finding the minimal value. The software for the above method has been implemented in the S language, based on its built-in functions for density estimation. Tools for visualisation of data have been included. One or more gels can be simultaneously plotted on the screen with bands assigned to different variants depicted in different colours. Modifications of the data can be interactively performed on screen. The software has been successfully used for several fingerprinting studies of maize inbreds.

Introduction

The recent years showed an important development of the new class of genetic markers which are based on the variation in the length of DNA fragments digested with the restriction endonucleases and which are called RFLP (Restriction Fragment Length

Polymorphism). Their application to plant breeding can be broadly divided into two categories: genetic fingerprinting and marker assisted selection. The genetic fingerprinting can be used to identify genotypes, to measure the level of heterozygosity/homozygosity, and to asses the relationship among inbreds, hybrids and populations. The marker assisted selection covers, among other things, backcross monitoring, identification of QTLs, and early selection of genotypes with desirable traits. The genetic fingerprinting is closely related to the issues of breeders' rights, essential derivation, and intellectual property, which are of the foremost interest to the seed companies.

The major advantages of the RFLP markers are that they are virtually unlimited in number, are not affected by the environment or the developmental state of the plant since they reveal differences at the DNA level, and can be characterised using a small amount of material obtained from seedlings.

The routine fingerprinting studies can involve several hundreds of genotypes that have to be analysed with hundred or more probe-enzyme combinations. For practical reasons, the genotypes will have to be disposed over several blots, each comprising two or more tiers (the width of the blot is usually about 30 lanes). In each lane we will observe one or more fragments ("bands") on the autoradiography (or no fragments at all), defining the RFLP profile of the genotype. The molecular weight of each fragment is computed from the migration distance of the fragment and the migration distances of standard fragments (of known molecular weights). The RFLP data should theoretically be in the form of a small number of discrete bands called variants (which are often called alleles, although they can not be considered as alleles in strict genetic sense). Since the observed data are subject to different sources of noise (gel distortion, measurement errors), they form a set of more or less continuous variables. The gel electrophoresis (autoradiograms of Southern blots) data are usually read into a computer by an image analysis system, which automatically translates migration distances to molecular weights, and can perform some corrections, but only for one blot at a time (or one tier at a time). Therefore we need a scoring procedure in order to assign a discrete value to each band. The "by eye" scoring can be easily performed when only a small number of genotypes is studied, but for a huge study involving several blots for each probe-enzyme combination it is more or less impossible (except for the probes with very simple patterns, but this usually means with low polymorphism and thus not very informative).

Method

As was previously stated, one expects the RFLP to be in the form of a small number of discrete bands. The observed data form a set of relatively continuous values because of the different distortions and errors occurring during the measurement process. An example of what the observed data may look like is shown in Figure 1. Figure 2 shows the same data after interpretation. This artificial example assumes that there are twenty genotypes studied, that there are five different "alleles" corresponding to the fragments of molecular weights of 1 to 5 kb, that there are four genotypes having each of the "alleles", and, to make things even nicer, that the genotypes are grouped in the lanes by their "alleles". Our aim was to construct an automatic smoothing procedure that will convert the observed data into interpreted data, which could readily be used for different computations (of distances, for example), required in the fingerprinting studies.

The approach we used is based on the non-parametric estimation of probability density function (Wegman 1972, Silverman 1986). The observed molecular weights were considered as a random variable whose probability density function we want to estimate. Figures 3 and 4 show the density functions of the molecular weights for the example data from Figures 1 and 2, respectively. The procedure we used is based on the density function in S (Becker *et al.* 1988). This is a kernel estimate. For each observed data value (x), the window is centred on that value and the heights of the window at each datapoint are summed. This sum, after a normalisation, is the corresponding function value (y) in the output.

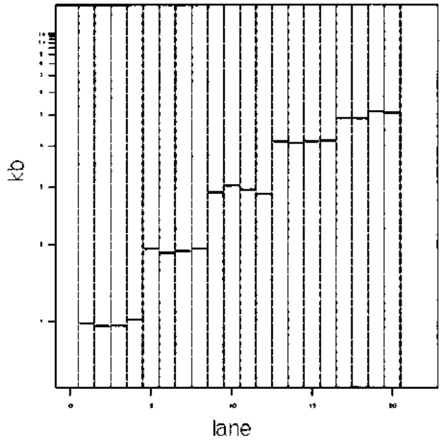


Figure 1. Example of observed data

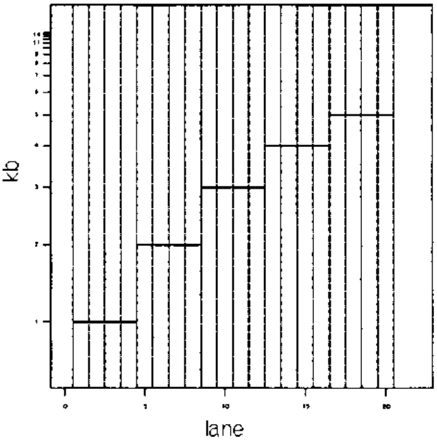


Figure 2. The example after interpretation

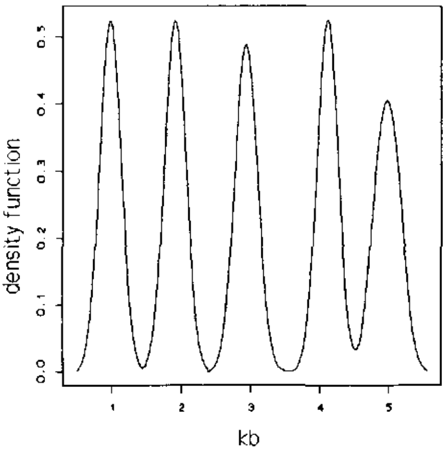


Figure 3. Density function of the molecular weights in Figure 1

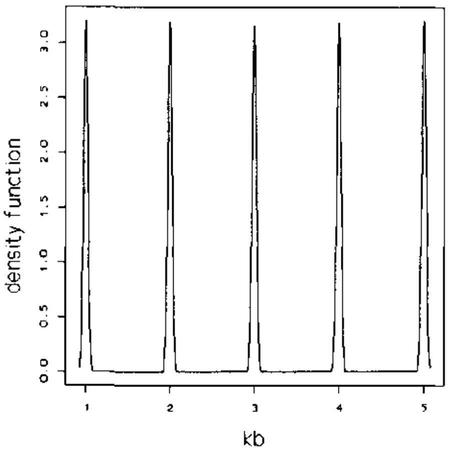


Figure 4. Density function of the molecular weights in Figure 2

Density estimation is essentially a smoothing operation. The key parameter is the choice of the window width (Silverman 1978, 1982). Inevitably there is a trade-off between bias in the estimate and the estimate's accuracy: wide windows will produce smooth estimates that may hide local features of the density. On the other hand, narrow windows may yield density estimates that model noise, not pattern. We found that in most cases the default setting of the window width gives reasonable density function estimates. The default is the width of a histogram bar which is determined by log2(number of data points) + 1 bars to cover the range of data values.

Once the estimate of density function is obtained, we look for the peaks of the function, which will give us the set of possible discrete values ("alleles") for the probe-enzyme combination being analysed. A peak (local maximum) is defined as an element in a sequence which is greater than all other elements within a window of specified width centred at that element. With the "alleles" defined, we must yet assign each observed data value to its appropriate "allele". This can be easily achieved by computing the distances between the observed data and all the "alleles", and assigning the band to the nearest possible discrete value. The distances are simply the differences in molecular weight between the observed data value and the "theoretical" molecular weight of the peak.

Software

Choice of the development tool

In order to implement the method described above, we decided to use the S-Plus statistical package (StatSci 1993), which is a commercially enhanced and fully supported release of the AT&T Bell Laboratories' S language (Becker *et al.* 1988). The reason for our decision is that S is both a statistical/graphical package and a very powerful programming language for creating new tools. The software offers a wide range of built-in statistical procedures, multiple active graphics windows, interactive graphics input and point identification using a mouse, and a very high level, structured, object-oriented language supporting classes, methods and inheritance.

The two major potential drawbacks of developing application software using a high-level language like S rather than using traditional programming languages like FORTRAN or C, are the risk of poor performance and the necessity to have a copy of S-Plus software for each user, both due to the fact that S is an interpreter and not a compiler. Since our application is single-user (to be used in our laboratory only), the second issue was not an obstacle. Concerning the performance, we needed a fast system, where the computer will be waiting for the user and not the inverse. The only bottleneck we encountered was assigning each observed data point to the nearest peak, which was solved by rewriting this small portion of the programme in the C programming language, and linking this module into S-Plus. (We could have done it in FORTRAN, too, since both C and FORTRAN code can easily be interfaced with S-Plus).

Features

The main feature of this software is, of course, automatic detection of the "alleles" and assignment of the observed data to the computed peaks. This tool was coupled with the visualisation of the gels, exploiting the graphical possibilities of the S-Plus software. One or more gels can be simultaneously plotted on the screen, giving an exact reproduction of the autoradiography, with bands assigned to different variants depicted in different colours. There are normally three open windows on the screen: one where all the blots are plotted, the second one where a selected tier can be examined in more detail and/or modified, and a third one displaying the "alleles". When there is doubt about the validity of the automatic scoring procedure, the user can interactively change the assignments by simply clicking with a mouse on the bands in question. Data that were wrongly recorded (radioactivity spots read in as bands) can also be corrected (deleted) this way. The bands that were not read in or were discarded by error when the gel was scanned can be interactively added with a mouse. The images of all tiers can be overlaid to give a general overview of the bands.

Other useful tools are also available: one can plot the estimated density function, compute the "allele" frequencies, compute the PIC (polymorphism information content) value of the probe-enzyme combination (Anderson *et al.* 1993), and perform a hierarchical cluster analysis (and plot a dendrogram) of the observed molecular weights. Everything is bundled-up in a user friendly, menu driven environment. No knowledge of S language is needed to use the software: all actions are accomplished by clicking the appropriate mouse button.

Short historical overview

The first version of the application was developed in 1992 using S-Plus for DOS release 2.0. It was tested by analysing a pilot study involving 17 maize (*Zea mays* L.) inbred lines evaluated for 105 probe-enzyme combinations. The results were compared with the results obtained by hand scoring, and, the conclusions being very favourable for the automated scoring procedure, it was decided to further enhance the application. The second version of our application was implemented in S-Plus for MS Windows release 3.1, which became available in 1993. Besides improving user interface and enhancing data visualisation, we added the direct access to an external (dBASE-compatible) database were all the data are stored. The data for each probe-enzyme combination are automatically retrieved from the database when needed, and the database is updated after the scoring has been done.

We are currently porting the application to the UNIX environment, where it will be interfaced with the SQL-based relational database engine. Some minor modification will be made to the software to make it more robust and the user interface will be further enhanced based on the dialogue building facility introduced in the release 3.2 of S-Plus for UNIX.

Discussion and conclusions

The software described above has been successfully used in three large scale fingerprinting studies of maize inbred lines. These studies involved 200, 150, and 250 inbred lines that were probed with, respectively, 90, 300, and 90 probe-enzyme combinations. Since the standard protocol used in our laboratory is two-tier blots with 30 lanes, up to 10 tiers had to be compared simultaneously for a given probe-enzyme combination (allowing 5 extra lanes per tier for molecular weight standard and/or standard inbred lines). On the other hand, up to 45000 bands per study had to be scored. Since, for some probe-enzyme combinations, up to 12 different "alleles" and 25 different profiles were found, it is clear that this kind of analysis would be impossible to conduct by hand scoring. The "alleles" found for a given probe-enzyme combination in different studies had very close, and often exactly the same, molecular weights. We can, therefore, envisage to use the mean values of the observed molecular weights, that were assigned to a given "allele", as its "genuine" molecular weight, and store this information in the database for future use.

Our experience shows that this automated procedure allows to process very quickly and smoothly huge amounts of data. The data processed this way are ready to be plugged into any standard statistical package for further analysis (distance computations, cluster analysis, etc.). This procedure also allows to pool the results from several studies conducted at different periods in time and/or in different laboratories, thus preserving the investment in previous analyses.

References

Anderson, J.A., G.A. Churchill, J.E. Autrique, S.D. Tanksley & M.E. Sorrells, 1993. Optimizing parental selection for genetic linkage maps, Genome 36: 181-186.

Becker, R.A., J.M. Chambers & A.R. Wilks, 1988. The New S Language, Wadsworth, Pacific Grove, California.

Silverman, B.W., 1978. Choosing the window width when estimating a density, Biometrika 65: 1-11.

Silverman, B.W., 1982. Kernel Density Estimation using the Fast Fourier Transform, Applied Statistics 31: 93-99.

Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis, Chapman and Hall, London.

Statistical Sciences, Inc., 1993. S-PLUS Reference Manual, Seattle, Washington.

Wegman, E.J., 1972a. Nonparametric Probability Density Estimation: I. A Summary of Available Methods. Technometrics 14: 533-546.

Wegman, E.J., 1972b. Nonparametric Probability Density Estimation: II. A Comparison of Density Estimation Methods. J. Statist. Comput. Simul. 1: 225-245.

Mapping earliness genes in tomato (*Lycopersicon esculentum*)

Pim Lindhout[1], Sjaak van Heusden, Gerard Pet, Johan W. van Ooijen, Hans Sandbrink, Ruud Verkerk[2], Ria Vrielink & Pim Zabel[2], DLO-Centre for Plant Breeding and Reproduction Research, PO Box 16, NL-6700 AA Wageningen, [1]present address: Wageningen Agricultural University, Department of Plant Breeding, PO Box 386, NL-6700 JA Wageningen, [2]Wageningen Agricultural University, Department of Molecular Biology, Dreijenlaan 3, NL-6703 HA Wageningen, The Netherlands

Key words

earliness, *Lycopersicon esculentum*, quantitative trait locus, QTL mapping, RFLP map

Summary

A breeding line, named "IVT-KT$_1$", was developed by crossing and selection with regard to earliness. Among its ancestors were two wild relatives, *L. pimpinellifolium* and *L. parviflorum*. IVT-KT$_1$ flowered and set fruit one to four weeks earlier than other cultivars.

To identify QTLs for earliness, an F$_2$ population was obtained by crossing IVT-KT$_1$ with the late true breeding cultivar "Premier". Three loci were identified associated with earliness, one of which was mainly associated with time to flowering, another with fruit setting time and a third one with ripening time. Two of these loci were also associated with fruit size.

Introduction

Earliness is generally defined as the number of days from sowing to the appearance of the first ripe fruit (Kemble & Gardner 1992). Among tomato cultivars a large variation for earliness exists (Baggett & Frazier 1978, Nieuwhof *et al.* 1986). Studies of the genetics of earliness in progenies obtained from intraspecific as well as interspecific crosses have indicated that dominance plays an important role in earliness (Banerjee & Kalloo 1989, Kemble & Gardner 1992). A negative correlation between earliness and fruit size is likely to be due to pleiotropic effects as breeding activities have not resulted in the release of early and large fruiting cultivars (Banerjee & Kalloo 1989).

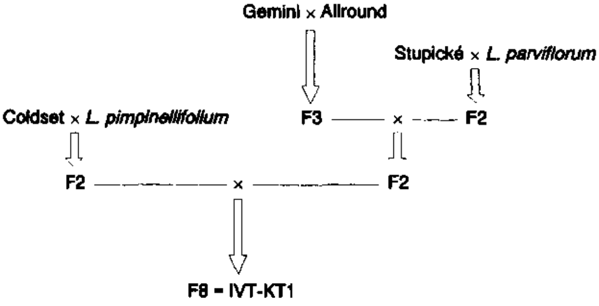
At CPRO-DLO, a breeding programme has been carried out to introduce earliness from *L. pimpinellifolium* into the cultivated tomato; to improve the plant vigour, *L. parviflorum* was included as a progenitor (Figure 1). One of the resulting early breeding lines, named IVT-KT_1, flowered and set fruit some weeks earlier than conventional cultivars. The earliness of IVT-KT_1 was also associated with a small fruit size.

Molecular markers provide a tool for mapping genes involved in quantitative traits (QTL), and as such they can be used to investigate whether correlated traits are affected by the same gene with pleiotropic effects, or by separate, linked genes. For tomato a detailed RFLP map is available (Tanksley *et al.* 1992). The present study aims at mapping genes involved in earliness in tomato and at studying the effects of these genes on components of earliness, especially on fruit size. For this purpose an F_2 population was made by crossing IVT-KT_1 with the late true breeding cultivar Premier. This F_2 population was analyzed for earliness and related components and subjected to RFLP analysis.

Polymorphisms between IVT-KT1 and Premier

To reveal polymorphisms 206 TG probes were hybridized onto blots with DNA of the two parents, digested with six restriction enzymes. Forty-seven probes showed a polymorphism, often with more than one restriction enzyme. The RFLPs were present in tight linkage groups (Figure 2). For example, of the 40 probes, known to map on chromosome 1, only two revealed a polymorphism. No RFLP was found for chromosome 7, whereas ten of the 17 probes for chromosome 3 revealed polymorphisms. All RFLP alleles of IVT-KT_1 that were different from those of Premier were indistinguishable from those of its ancestors *L. pimpinellifolium* and/or *L. parviflorum*, suggesting that linked chromosome fragments originate from these ancestors.

Figure 1. The ancestry of the early breeding line IVT-KT_1. The names refer to *L. esculentum* cultivars or accessions of related *Lycopersicon* species. During subsequent selfings selections were carried out for early fruit set at low light and low temperature conditions



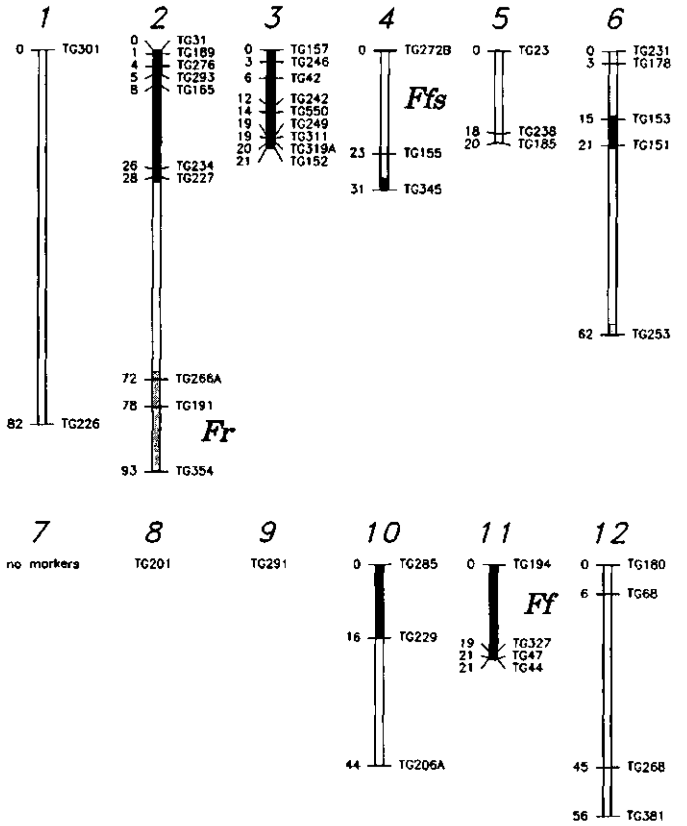


Figure 2. The map position of RFLPs and earliness QTLs in the F2 between IVT-KT$_1$ and Premier. Black fragments originate from *L.parviflorum* and hatched fragments from *L.pimpinellifolium*. The other markers were identical in the two wild relatives

Variation in earliness

In the winter and early spring of 1990/1991, 690 plants of the F_2 (IVT-KT$_1$ x Premier) and 12 control genotypes were tested for characters related to earliness. The results are shown in Table 1. IVT-KT$_1$ was 23 to 33 days earlier in Flowering, Fruit setting and Ripening than Premier (see Table 1 for the definition of the traits). IVT-KT$_1$ also had a smaller number of leaves under the first inflorescence (LeavesNo) and a smaller Fruit size (*cf.*, Nieuwhof *et al.* 1987). The F_1 showed overdominance for most of the traits. For all traits (except Fruit size) the F_2 means were closer to the mid-parent value than the F_1-mean. Among the ancestors of IVT-KT$_1$, a large variation in earliness was found. The cultivar Stupické and *L. pimpinellifolium* were about as early as IVT-KT$_1$, while

Table 1. Plant characters related to earliness of tomato genotypes and some wild relatives. Flowering = time after sowing to appearance of first open flower at first inflorescence (days), Fruit setting = time after sowing to appearance of first fruit with diameter larger than 3 cm (days), LeavesNo = number of true leaves under the first inflorescence, Earliness = time after sowing to first ripe fruit (days), Ripening = time between Fruit setting and first ripe fruit (days), Fruit size = average fruit size of harvested early fruits (g)

Genotype	Flowering	Fruit setting	LeavesNo	Ripening	Earliness	Fruit size
1. IVT-KT$_1$	91	101	9.4	47	148	20
2. Premier	114	126	11.4	55	181	51
3. F_1	84	93	8.3	56	150	35
4. F_2	88	101	8.9	53	155	38
5. Coldset	108	122	9.8	49	171	66
6. *L. pimpinellifolium*	86	100	11.3	53	153	5
7. Gemini	89	97	10.4	54	151	47
8. Allround	92	103	10.3	59	162	66
9. Stupické	86	96	10.1	50	146	40
10. *L. parviflorum*	87	105	10.3	58	163	1
11. Liberto	87	97	9.8	61	158	70
12. Moneymaker	91	102	10.2	59	161	60
13. Rapide	90	100	9.4	60	160	7
Heritability:	0.92	0.84	0.72	0.85	0.62	0.45

Coldset and the wild relative *L. parviflorum* were late. The cultivars Gemini, Moneymaker, Liberto and Rapide performed intermediate for nearly all characters except for fruit size, which was larger than average.

The RFLP linkage map of the F_2

The 145 earliest and 147 latest F_2 plants were subjected to RFLP analysis by using 45 probes. The RFLP linkage map was obtained using JoinMap (Stam 1993). Since the polymorphic probes were not equally distributed over the tomato genome, it was impossible to generate an RFLP map, that covered the whole genome (Figure 2). No chromosome 7 specific RFLP was found. Three markers - TG157, TG151 and TG268 - mapped on chromosomes 3, 6 and 12 respectively, while in the *L. esculentum* × *L. pennellii* map they have been located on chromosomes 1, 2 and 4, respectively (Tanksley *et al.* 1992). Generally, where the distances between the markers of the present map could be compared with those of the *L. esculentum* × *L. pennellii* map, they were similar.

Mapping QTLs for earliness and related characters

The map positions of genes involved in earliness and related characters were estimated on the basis of the plants evaluations and RFLP data of 292 selected F_2 plants. We

Table 2. Pleiotropic effects of QTLs for earliness. The figures represent the explained variance per locus. Chromosomes that did not harbour significant QTLs, were omitted. The totals indicate the total explained variance assuming that the effects are additive

Trait		Chromosome			Total
		2	4	11	
	RFLP marker:	TG354	TG155	TG194	
	Locus name:	*Fr*	*Ffs*	*Ff*	
Flowering		-[1]	-	23	23
Fruit setting		-	9	17	26
LeavesNo		-	7	44	51
Ripening		12	-	-	12
Earliness		4	9	16	29
Fruit size		21	-	14	35

[1] Nonsignificant effects are indicated with '-'

applied the interval mapping procedure (Lander & Botstein 1989) using the computer program MapQTL developed at CPRO-DLO. A significance threshold of 3.7 LOD was employed (Van Ooijen 1992). For Earliness three QTLs were detected on chromosomes 2, 4 and 11 (Table 2, Figure 2). They accounted for 4, 9 and 16% of the total phenotypic variance, respectively. IVT-KT$_1$ alleles on the loci of chromosomes 2, 4 and 11 enhanced earliness. Additionally, the Earliness QTL on chromosome 2 was associated with ripening time and fruit size, the chromosome 4 QTL with fruit setting time, and the chromosome 11 QTL with time to flowering, fruit setting time and fruit size. Accordingly, we denote the Earliness QTLs on chromosome 2, 4 and 11 by *Fr*, *Ffs* and *Ff* respectively, (*fast ripening, fast fruit setting* and *fast flowering*). The heterozygote *Ffff* was as early as *FfFf*, indicating a complete dominant inheritance; the other loci showed intermediate inheritance (data not shown). For all other characters significant QTLs were found, with an explained variance per ranging from 4 to 44%.

Confirmation of the QTL detection with selected F$_3$ lines

To confirm the map position and the quantitative effects of the earliness loci, 22 F$_2$ plants, homozygous for the alternative alleles at *Fr* and *Ff*, were selected for progeny testing. In the winter and early spring of 1991/1992, 16 plants of each selected F$_3$ line were grown and evaluated for earliness and related characters (Table 3). Seven F$_3$ lines homozygous for *Fr* and *Ff* (IVT-KT$_1$ allele) were nearly as early as IVT-KT$_1$ whereas six F$_3$ lines homozygous for the same loci (Premier allele) were as late as Premier. The difference between these two groups was 23 days covering most of the 29 days difference between the parents (Table 3). The F$_3$ lines homozygous for one locus

Table 3. The mean earliness of F_3 lines and their parents

Genotype	No. of plants	Chr. 2[1]	Chr. 11[2]	Earliness (days)
IVT-KT_1	32	a[3]	a	123
Premier	32	b	b	153
F_1	32	h	h	132
F_3 lines	7×16	a	a	128
F_3 lines	7×16	a	b	143
F_3 lines	2×16	b	a	140
F_3 lines	6×16	b	b	151
Liberto	16			144
Rapide	16			146

[1] Chromosome 2 was represented by TG266, TG191 and TG354. [2] Chromosome 11 was represented by TG194, TG327, TG44 and TG47. [3] a, h, or b: homozygous for IVT-KT_1 chromosome fragments, heterozygous, or homozygous for Premier chromosome fragments, respectively

(IVT-KT_1 allele) and homozygous for the other locus (Premier allele) showed intermediate earliness.

Fine mapping the *Ff* locus

To obtain a better estimation of the position of *Ff* on chromosome 11, we tested more RFLP markers known to map in the region of interest between TG194 and TG44. Furthermore, Bulked Segregant Analysis (BSA) was carried out to find additional RAPD markers in this area (Michelmore *et al.* 1991, Giovanni *et al.* 1991). This resulted in a more saturated map (Figure 3). With this map we were able to study the co-segregation of markers and Earliness in specific F_3-lines, which gave us the indication that two, instead of one, QTLs segregated on this chromosome (Figure 4).

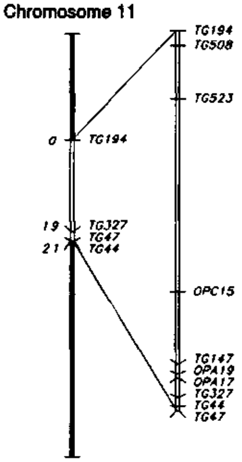
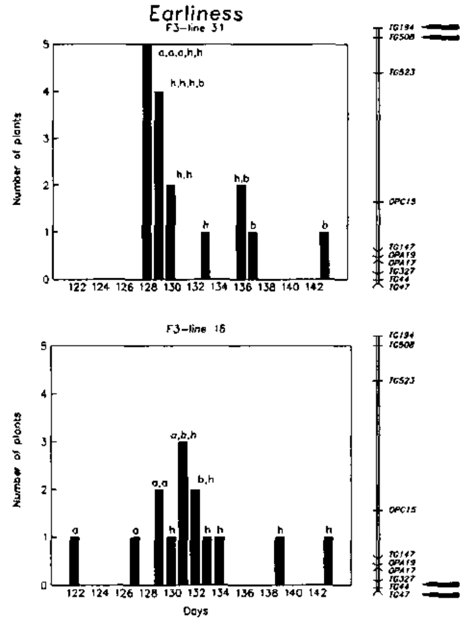


Figure 3. Integrated map showing RFLP and RAPD markers on chromosome 11 in the region of *Ff*. The outer bars indicate the position of the outer markers in the map by Tanksley *et al.* (1992)

Breeding perspectives

In the present study, it was shown that three loci were involved in the segregation for earliness in the F_2 of the cross IVT-KT_1 × Premier. Together they accounted for 29% of the total phenotypic variation for earliness. Compared to the estimated heritability of 62%, these results indicate that about half the total genetic variation for earliness is explained by these three QTLs, so other QTLs for earliness may not have

Figure 4. Co-segregation of earliness with markers on chromosome 11; the number of plants of F_3-lines 31 (top panel) and 16 (bottom panel) in relation to Earliness. **Top panel:** F_2-parent no. 31 is heterozygous for TG194 and TG508 and homozygous IVT-KT_1 alleles for the rest of the markers. The genotypes of the F_3-plants are indicated with **a**, **h**, or **b**: homozygous IVT-KT_1 alleles, heterozygous, or homozygous Premier alleles for TG194 and TG508, respectively. **Bottom panel:** F_2-parent no. 16 is heterozygous for TG44 and TG47 and homozygous IVT-KT_1 for the rest of the markers. The genotypes of the F_3-plants are indicated with **a**, **h**, or **b**: homozygous IVT-KT_1 alleles, heterozygous, homozygous Premier alleles for TG44 and TG47



been identified. Among the *L. esculentum* ancestors of the breeding line IVT-KT_1, were early lines, some of them as early as IVT-KT_1 (Figure 1, Table 1). The presence or absence of genes from these ancestors in IVT-KT_1 could not be demonstrated due to lack of polymorphisms between these lines and Premier.

In view of the polymorphisms detected in the present study and the known association of a high level of genetic variation between *Lycopersicon* species but not within *L. esculentum*, the three chromosome fragments carrying a gene for earliness are likely to originate from the wild ancestor species *L. parviflorum* and *L. pimpinellifolium* (Miller & Tanksley 1990, Van der Beek *et al.* 1992).

L. pimpinellifolium has been used before as parent in breeding for earliness (Baggett & Frazier 1978). On the basis of the vigourousness of the F_1 with the cultivated tomato, *L. parviflorum* had been included as ancestor of IVT-KT_1 to increase plant vigour. Remarkably, in the present study, the *L. parviflorum* ancestor was also early and donated two of the three earliness loci to IVT-KT_1.

The present study also aimed at getting more insight in the association between earliness and fruit size. It was shown that each earliness locus had a different relative contribution to earliness and fruit size. The *Fr* locus on chromosome 2 accounted for 4%

of the earliness variation but 21% of the fruit size variation whereas the *Ff* locus on chromosome 11 explained 16% of the variation for earliness and 14% for fruit size. Finally, the *Ffs* locus explained 9% of the earliness variation but had no significant effect on fruit size. These data allow breeders to identify those genes which can be used in their breeding lines. For example, although *Ffs* may not have the largest effect on earliness, its lack of effect on fruit size may render this locus more valuable than other QTLs. Thus the breeder now has the opportunity to predict more precisely the results of his breeding efforts.

Acknowledgements

This research was sponsored by eight Dutch breeding companies and the Netherlands Ministry of Economic Affairs. Ad Hermsen and Harm Wiegersma are acknowledged for nursing the plants. A more extensive version of this manuscript was originally published in: Proceedings of the Tomato Biotechnology Symposium, Hille *et al.* (Eds.), Kluwer Academic Publishers, Dordrecht, The Netherlands.

References

Baggett, J.R. & W.A. Frazier, 1978. Oregon cherry tomato. HortScience 13: 598.

Banerjee, M.K. & Kalloo, 1989. The inheritance of earliness and fruit weight in crosses between cultivated tomatoes and two wild species of *Lycopersicon*. Plant Breeding 102: 148-152.

Giovanonni J.J., R.A. Wing, M.W. Ganal & S.D. Tanksley, 1991. Isolation of molecular markers from specific chromosomal intervals using DNA pools from existing mapping populations. Nucl. Acids Res. 19: 6553-6558.

Kemble, J.M. & R.G. Gardner, 1992. Inheritance of shortened fruit maturation in the cherry tomato Cornell 871213-1 and its relation to fruit size and other components of earliness. Journal of the American Society for Horticultural Science 117: 646-650.

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

Michelmore R.W., I. Paran & R.V. Kesseli, 1991. Identification of markers linked to disease resistance genes by Bulked Segregant Analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. Proc. Natl. Acad. Sci. 88: 9828-9832.

Miller, J.C. & S.D. Tanksley, 1990. RFLP analysis of phylogenetic relationships and genetic variation in the genus *Lycopersicon*. Theoretical and Applied Genetics 80: 437-448.

Nieuwhof, M., G. Pet & F. Garretsen, 1987. Inheritance of characters determining growth and development of tomato (*Lycopersicon esculentum* Mill.) under low energy conditions. Euphytica 36: 205-213.

Stam, P., 1993. JoinMap: a computer package to construct integrated genetic linkage maps. Plant Journal 3: 739-744.

Tanksley S.D., M.W. Ganal, J.P. Prince, M.C. Devicente, M.W. Bonierbale, P. Broun, T.M. Fulton, J.J. Giovanni, S. Grandillo, G.B. Martin, R. Messeguer, J.C. Miller, L. Miller A.H. Paterson, O. Pineda, M.S. Roder, R.A. Wing, W. Wu & N.D. Young, 1992. High density molecular linkage maps of the tomato and potato genomes. Genetics 1324: 1141-1160.

Van der Beek, J.G., R. Verkerk, P. Zabel & P. Lindhout, 1992. Mapping strategy for resistance genes in tomato based on RFLPs between cultivars: *Cf9* (resistance to *Cladosporium fulvum*) on chromosome 1. Theoretical and Applied Genetics 84: 106-112.

Van Ooijen, J.W., 1992. Accuracy of mapping quantitative trait loci in autogamous species. Theoretical and Applied Genetics 84: 803-811.

QTL mapping in a full-sib family of an outcrossing species

C. Maliepaard & J.W. van Ooijen, DLO-Centre for Plant Breeding and Reproduction Research (CPRO-DLO), P.O. Box 16, 6700 AA Wageningen, The Netherlands

Key words

allogamous species, full-sib family, QTL mapping

Summary

Methods for mapping quantitative trait loci (QTLs) using molecular markers are widely applied in autogamous plant species. One of the methods often used is the interval mapping method. In this method the information of flanking markers is used for the detection of a QTL in a marker interval. In BC_1 and F_2 populations, the markers flanking a QTL provide the maximum amount of information. In a full-sib (FS) family of an outcrossing species, different marker intervals contain different amounts of information about a QTL. The markers flanking a QTL are not necessarily the most informative ones. As a consequence, the graph of the LOD score may show discontinuities between marker intervals and a QTL present in one interval may be mapped in a neighbouring, more informative interval.

These difficulties can be overcome by the simultaneous use of all available marker information. Simulation results show that neighbouring markers may compensate missing information. Discontinuities in the graph of the test statistic can be removed and QTLs can be detected which would not have been found if only flanking markers had been considered.

Introduction

With the availability of linkage maps with large numbers of molecular markers for several animal and plant species, methods for mapping quantitative trait loci (QTLs) have become an important tool in the genetic analysis of quantitatively inherited characters. For plants, most attention has been focused on segregating progeny of crosses involving fully homozygous parents, *i.e.*, first generation backcross (BC_1) or F_2

populations. For these population types, Lander & Botstein (1989) developed the so-called interval mapping method that uses the genotypic data of two flanking markers and the quantitative trait values for calculating the likelihood for the presence of a QTL for every position on the linkage map. This interval mapping method can be extended to outcrossing species, although several difficulties arise compared to autogamous species, especially in the case of quantitative traits. First, we will discuss some of these difficulties.

The major problem in QTL mapping in outcrossing species concerns the fact that two to four alleles may be involved in the segregation within the FS family, and this may vary between markers. The different types of segregation are listed in Table 1. In the BC_1-types of segregation only one of the parents can provide information with respect to linked QTLs, whereas in the other types the meioses of both parents are informative. In the case of markers with an F_2 segregation type it is unknown for the heterozygous progeny which allele was derived from which parent. The ideal markers for QTL mapping are those segregating for 3 or 4 alleles, essentially because all parental gametes can be retraced unambiguously for all genotypes of the progeny. Screening of the parents may enable the selection of only these 3- or 4-allele markers. Currently, however, the number of such markers is insufficient, and often the 2-allele markers must be employed.

A second problem is that, in general, the linkage phase between loci in both parents is not known in advance. This information has to be deduced from the genotypic data in the linkage analysis. When dominant markers are involved this can sometimes be problematic.

A last and more important problem is the number of alleles and the type of segregation of the QTL in a FS family. First, parents in a cross may be fixed for alternative QTL alleles (qq versus QQ). As a consequence, no segregation will occur and these QTLs cannot be mapped, even though they contribute to the difference between the

Table 1. Segregation types of codominant loci in outcrossing species. Different characters indicate different alleles of a locus

$P_1 \times P_2$	Description
aa × ab	BC_1-type, 2 alleles, P_2 heterozygous
ab × aa	BC_1-type, 2 alleles, P_1 heterozygous
aa × bc	BC_1-type, 3 alleles, P_2 heterozygous
bc × aa	BC_1-type, 3 alleles, P_1 heterozygous
ab × ab	F_2-type, 2 alleles
ab × ac	3 alleles
ab × cd	4 alleles

parents. Second, when segregation occurs, the possibility of more than two QTL alleles should be considered. At least for certain loci, multiple alleles are known to exist in outcrossing species, *e.g.*, self-incompatibility genes and molecular markers. With respect to QTLs, three alleles were detected for a QTL segregating in a single FS family of diploid potato (Van Eck *et al.* 1994). Because the number of QTL alleles is not known in advance, a general QTL mapping approach for an outcrossing species should take into account the segregation of four alleles. Situations with less than four segregating alleles can be considered as simplifications.

As a result of these difficulties, genetic analysis of quantitative traits in outcrossing plant species has been limited. Recently, Knott & Haley (1992) developed a maximum likelihood method for mapping QTLs in multiple FS families, a strategy necessary for instance in pig breeding, where only small numbers of progeny per family are available. In comparison with mapping in a single FS family, some additional problems have to be taken into account. These are (1) genetic and environmental variation between families, (2) fixation of QTLs in some of the families, and (3) differences in marker segregation types *between* families. Knott & Haley found discontinuities in the graph of the LOD score, which were caused by problems (2) and (3), but also by differences in segregation type *between* markers *within* a family. The latter is also of importance in QTL mapping using only a single FS family. The result of the discontinuities can be an incorrect localization of a QTL. In order to resolve this problem Knott & Haley proposed to use the information of several or all available markers on a chromosome. Haley *et al.* (1994) showed that the simultaneous use of multiple markers from a linkage group in a regression mapping method can remove these discontinuities, and, moreover, increase the power of QTL detection.

We developed a maximum likelihood method using all markers from a linkage group, analogous to the regression method of Haley *et al.* (1994). This method is called the "all-markers" mapping method. This paper shows in two examples how the interval mapping method, extended for allogamous species, is affected when flanking markers are not fully informative with respect to a QTL. The gain of using the all-markers mapping method is demonstrated.

Interval mapping for allogamous species

The QTL mapping procedure as described by Lander & Botstein (1989) and more extensively worked out by Van Ooijen (1992) for autogamous species, is adopted here for a FS family of an outcrossing species.

For a QTL four segregating alleles are assumed. Indicating QTL alleles derived from P_1 with *1* and *2* and QTL alleles from P_2 with *3* and *4*, the segregation can be represented as *12* × *34*, and the four resulting QTL genotypes in the progeny as *13*, *14*, *23* and *24*. Four normal probability distributions with means μ_{13}, μ_{14}, μ_{23}, μ_{24} and equal residual variance σ_r^2 are assumed for the QTL genotypes. At a map position between two markers, determined by recombination frequencies r_a and r_b, the mixture probability density function (pdf) for an individual i with marker genotype m_i and phenotypic value y_i of the quantitative trait is specified:

$$f(y_i|m_i;r_a) = \pi_{m13}f_{m13}(y_i) + \pi_{m14}f_{m14}(y_i) + \pi_{m23}f_{m23}(y_i) + \pi_{m24}f_{m24}(y_i) \ ,$$

where π_{mq} is the probability of a QTL genotype ($q \in \{13, 14, 23, 24\}$) given marker genotype m, and $f_q\ (y_i)$ is a normal pdf with mean μ_q and variance σ_r^2. The mixture model is tested against the model of the null hypothesis where no QTL is segregating, H_0: $\mu_{13} = \mu_{14} = \mu_{23} = \mu_{24}$. The test statistic is the LOD score (Lander & Botstein 1989).

In a BC_1 or an F_2 population the flanking markers provide the maximum amount of information with respect to a QTL, unless there are many missing values or a marker is dominant. In that case a neighbouring marker may provide additional information. In a FS family of an outcrossing species, markers segregating for three or four alleles also provide the maximum amount of information with respect to a QTL in the interval. Markers segregating for only two alleles cannot always distinguish between the QTL genotypes. If the probabilities of two QTL genotypes, π_{mq}, are equal within all genotype classes of a pair of flanking markers, these markers cannot distinguish between these two QTL genotypes. It will be clear that a pair of flanking markers with an aa × ab segregation type will not allow the distinction of all four genotypes. Only two distributions can be fitted, since $\pi_{13} = \pi_{23}$ and $\pi_{14} = \pi_{24}$ for all marker genotype classes. Similarly, two markers with an F_2 type of segregation and in coupling phase cannot distinguish between two of the QTL genotypes. Table 2 illustrates this. The probabilities π_{14} and π_{23} are equal within all marker genotype classes. As a result the maximum likelihood estimates for μ_{14} and μ_{23} are equal, so that in fact only three distributions with expected means μ_{13}, μ_{24} and $(\mu_{14} + \mu_{23})/2$ are fitted.

All-markers mapping

If markers flanking a QTL have more informative neighbours, these can contribute additional information with respect to a QTL. Even if those neighbouring markers are

Table 2. Probabilities (multiplied by 4) of QTL genotypes in a full-sib family of an outcrossing species. Segregation type (the haplotypes are separated by ' / '): a*1*a / b*2*b × a*3*a / b*4*b, r = recombination fraction between first marker locus and QTL, v = recombination fraction between second marker locus and QTL, $s = 1-r$, $w = 1-v$

Marker genotype		QTL genotype			
Left	Right	*13*	*14*	*23*	*24*
aa	aa	s^2w^2	$rsvw$	$rsvw$	r^2v^2
	ab	$2s^2vw$	$rs(v^2+w^2)$	$rs(v^2+w^2)$	$2r^2vw$
	bb	s^2v^2	$rsvw$	$rsvw$	r^2w^2
ab	aa	$2rsw^2$	$(r^2+s^2)vw$	$(r^2+s^2)vw$	$2rsv^2$
	ab	$4rsvw$	$(r^2+s^2)(v^2+w^2)$	$(r^2+s^2)(v^2+w^2)$	$4rsvw$
	bb	$2rsv^2$	$(r^2+s^2)vw$	$(r^2+s^2)vw$	$2rsw^2$
bb	aa	r^2w^2	$rsvw$	$rsvw$	s^2v^2
	ab	$2r^2vw$	$rs(v^2+w^2)$	$rs(v^2+w^2)$	$2s^2vw$
	bb	r^2v^2	$rsvw$	$rsvw$	s^2w^2

not fully informative themselves, they can still compensate a part of the missing information. We developed a QTL mapping algorithm that employs neighbouring marker information in case the information of the flanking markers is incomplete. Additional information is collected from other markers alongside the chromosome until the maximum is reached or we get to the last marker. The information for an individual in the progeny reaches a maximum when the parental contributions of marker alleles are unambiguous for any pair of markers at both sides of the map position tested.

Simulation study to compare interval mapping with all-markers mapping

In a small computer simulation study the effect of using all available marker information has been investigated. Two examples are used to illustrate the approach and the gain involved in the all-markers method compared to interval mapping using flanking markers only. In the examples a FS family of 200 progeny has been generated in each of 10 simulation runs. One chromosome of 60 cM map length with four markers at 0, 20, 40 and 60 cM and a QTL at 30 cM was simulated. The QTL has an F_2-type of segregation (Qq × Qq). Dominance is absent and the expected fraction of the variance explained by the QTL is 0.20 ($\mu_{qq} = 0$, $\mu_{qQ} = \mu_{Qq} = 1$, $\mu_{QQ} = 2$, $\sigma_r^2 = 2$).

Example 1

The markers at map positions 20 and 40 cM have a BC_1-type of segregation, both heterozygous in the male parent. The two outer markers segregate for four alleles. Schematically the cross can be represented as:

$$\frac{a\ a\ q\ a\ a}{b\ a\ Q\ a\ b} \times \frac{c\ a\ q\ a\ c}{d\ b\ Q\ b\ d} .$$

Figure 1 shows the results of interval mapping using only the flanking markers and of the all-markers method using all marker information. The markers flanking the QTL provide information with respect to the segregating QTL only in the heterozygous (male) parent, whereas any effect caused by the segregation of the QTL alleles in the female parent is obscured. The outer markers can contribute additional information even though they are at a larger map distance. Using their information results in a continuous LOD score graph, a correct positioning of the QTL (Figure 1) and better maximum likelihood estimates for the μs and σ_r^2. However, in the ideal situation, when all markers would segregate for four alleles, an even higher LOD score would have been obtained (Figure 1, dotted graph).

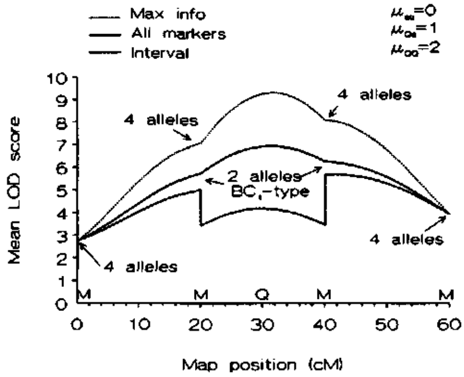


Figure 1. Mean LOD scores from 10 simulation runs, each one generating a FS population of size 200, with markers (M) at every 20 cM of a chromosome with a total map length of 60 cM. A QTL (Q) with an F_2-type of segregation (Qq × Qq) is present at 30 cM. The outer markers segregate for four alleles. The middle markers segregate for two alleles (aa × ab; dashed line: interval mapping, continuous line: all-markers mapping), or for four alleles (ab × cd, dotted line)

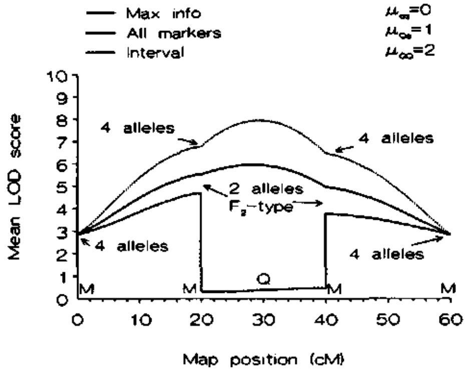
Example 2

The markers at 20 and 40 cM have an F_2-type of segregation (coupling phase in both parents). The outer markers segregate for four alleles. The QTL is in coupling phase with each of the flanking markers in only one of the parents. Schematically the cross can be represented as:

$$\frac{a\ a\ q\ a\ a}{b\ b\ Q\ b\ b} \times \frac{c\ a\ Q\ a\ c}{d\ b\ q\ b\ d} .$$

In this situation (Figure 2) no distinction can be made between the homozygous QTL genotypes, since their expected frequencies are equal within all marker genotype classes of the flanking markers (see Table 2: qq = *14* and QQ = *23*). Because of this and because the mean of the heterozygote equals the average of the means of the homozygotes ($\mu_{Qq} = (\mu_{qq} + \mu_{QQ})/2$), the QTL will not be detected when only flanking markers are used. The use of the additional information provided by the outer markers again results in a continuous LOD score graph which allows the correct positioning of the QTL and the estimation of the QTL effects.

Figure 2. Mean LOD scores from 10 simulation runs, each one generating a FS population of size 200, with markers (M) at every 20 cM of a chromosome with a total map length of 60 cM. A QTL (Q) with an F_2-type of segregation (Qq × Qq) is present at 30 cM. The outer markers segregate for four alleles. The middle markers segregate for two alleles, (ab × ab; dashed line: interval mapping, continuous line: all-markers mapping), or for four alleles (ab × cd, dotted line). The QTL is in coupling phase with the middle markers in one of the parents



It can be concluded that the problem of non-constant marker information in QTL mapping in a FS family can be solved by using all available marker information. Moreover, it enables the detection of QTLs which, with the use of flanking markers only, could remain undetected.

References

Haley, C.S., S.A. Knott & J. Elsen, 1994. Mapping quantitative trait loci in crosses between outbred lines using least squares. Genetics 136: 1195-1207.

Knott, S.A. & C.S. Haley, 1992. Maximum likelihood mapping of quantitative trait loci using full sib families. Genetics 132: 1211-1222.

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

Van Eck, H.J., J.M.E. Jacobs, P. Stam, J. Ton, W.J. Stiekema & E. Jacobsen, 1994. Multiple alleles for tuber shape in diploid potato detected by qualitative and quantitative genetic analysis using RFLPs. Genetics 137: 303-309.

Van Ooijen, J.W., 1992. Accuracy of mapping quantitative trait loci in autogamous species. Theor. Appl. Genet 84: 803-811.

Constructing confidence intervals for QTL location

B. Mangin, B. Goffinet & A. Rebaï, Institut National de le Recherche Agronomique, Station de Biométrie et d'Intelligence Artificielle, BP 27, 31326 Castanet-Tolosan Cedex, France

Key words
QTL position, confidence interval, similar statistics

Summary

We propose a method for the construction of confidence intervals for QTL location. This method, developed in a local asymptotic framework, leads to a likelihood ratio test based on statistics of which the asymptotic distribution does not depend on nuisance parameters, in particular the QTL effect. Using simulations, we compare this new confidence interval with the classic Lander & Botstein (1989) confidence interval and an empirical confidence interval proposed by Darvasi *et al.* (1993). The classic confidence interval can be biased for QTLs with a small effect. We show that the new confidence interval provides approximately correct coverage probabilities for almost all QTLs.

Introduction

Since Sax (1923) the literature contains many publications concerned with the detection of quantitative trait loci (QTL) using marker information. Estimation of the location of QTLs on a linkage map is possible using "interval mapping" procedures based on the maximum likelihood method (Lander & Botstein 1989). As pointed out by Darvasi *et al.* (1993), it is very important to consider confidence intervals for the locations of QTLs on the chromosomes. Conneally *et al.* (1985), in the field of linkage analysis, and Lander & Botstein (1989) proposed the use of a confidence interval based on the limiting χ^2 distribution of the likelihood ratio test used in the interval mapping procedure. In this paper, we propose a new confidence interval, compare it with the classic Lander & Botstein (1989) confidence interval and with the confidence interval proposed by Darvasi *et al.* (1993).

Model

We consider a backcross population of size n. A QTL is assumed to be present at position d on a chromosome of length L. The values of the trait considered follow a normal distribution with means μ_A and μ_B for the two QTL genotypes (A and B) present in the backcross population, with the same variance σ^2 for both genotypes. We will use $a = \mu_A - \mu_B$ for the QTL effect and $\mu = (\mu_A + \mu_B)/2$ for the grand mean.

Classic confidence interval

The classic confidence interval is based on the statistic $T(d_0)$, given by

$$T(d_0) = \sup_d LOD(d) - LOD(d_0) ,$$

where $LOD(d)$ is the ($\log_{10}$ based) likelihood ratio test statistic for position d. Its analytic expression is given by Lander & Botstein (1989).

To investigate the quality of this confidence interval, a simulation study was carried out for different values of the percentage of variance due to the QTL. Values of the percentage of variance were set equal to $100(a^2/4)/(a^2/4+\sigma^2)$. Results are given in Figure 1. It appears that for a QTL with a large effect the classic confidence interval is unbiased with markers at each 20 cM, but conservative with markers at each 5 cM. It is very biased for QTLs with a small effect, particularly in the case of a dense map.

The reason for this is that for a QTL with a small effect, $T(d_0)$ does not follow a $0.217\chi_1^2$ distribution under the null hypothesis as expected from classic asymptotic theory. Table 1 shows that 10%-quantiles obtained for the distribution of $T(d_0)$ are

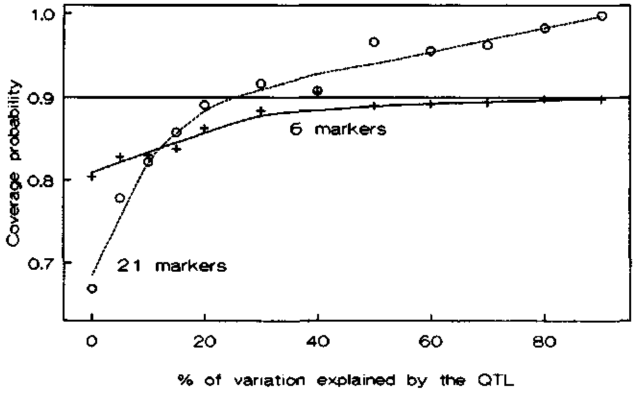


Figure 1. Empirical coverage probability for $T(d_0)$ over 1000 replications. In each simulation, a 200 backcross progeny is generated with a 100 cM chromosome, 6 or 21 markers equally spaced along the chromosome, $\mu = 0$, $\sigma^2 = 1$ and a QTL in the middle of the chromosome. The threshold used, is based on a $0.217\chi^2$ with 1 d.f. to ensure a 90% confidence interval

Table 1. 10%-quantiles of $0.217\chi_1^2$, $T(d_0)$ and the power of the LOD score test

	% Variance due to QTL	10%-Quantile	95% Confidence interval	Power
$0.217\chi_1^2$		0.59		
$T(d_0)$	2%	0.87	0.82-0.94	36%
	4%	0.83	0.77-0.91	63%
	10%	0.79	0.72-0.87	97%
	50%	0.59	0.56-0.66	100%

Empirical 10%-quantiles of $T(d_0)$ and the empirical power of the LOD score test for a type I error of 5% based on 1000 simulations. In each simulation, a backcross progeny of 200 individuals was generated with one 100 cM chromosome, 6 markers equally spaced at 20 cM, $\mu = 0$, $\sigma^2 = 1$ and a QTL in the middle of the chromosome

usually larger that the 10%-quantile of a $0.217\chi_1^2$ distribution. The difference depends on the percentage of variance due to the QTL, and the difference is large when the percentage of variance is small.

Constructing a "similar confidence interval"

In order to deal with QTLs with a small effect, Mangin *et al.* (1994) propose the use of a so-called "similar test", as described by Cox & Hinkley (1974). The basic idea is to find statistics of which the distribution does not depend on the nuisance parameter under the null hypothesis, *i.e.*, the QTL is located at d_0. Furthermore, one should work in a so-called "local asymptotic framework". This framework is used in asymptotic theory to obtain the power of maximum likelihood ratio tests, of which the asymptotic power is not trivially equal to 100%. It is the framework that should be used when dealing with QTLs which can be detected with powers ranging from 20% to 90% (Rebaï *et al.* 1993). Formally, in a local asymptotic framework, as the population size n tends to infinity, the QTL effect a is assumed to tend to 0 in such a way that $a\sqrt{n}$ converges to a finite constant δ.

Define $\mathbf{Z}(d_0)$ as the vector of components $Z_j(d_0)$ $(j = 1, 2, ..., J-1)$ given by

$$Z_j(d_0) = \frac{1}{\sqrt{\hat{\sigma}^2}}\left[\frac{S_j}{1-2r_{j,d_0}} - \frac{S_{j+1}}{1-2r_{j+1,d_0}}\right],$$

where r_{j,d_0} denotes the recombination frequency between marker j and a QTL located at d_0, $\hat{\sigma}^2$ is the classic estimate of the variance and S_j is the mean difference between the marker genotypes for marker j.

Proposition 1 (for a proof see Mangin *et al.* (1994)) shows that $\mathbf{Z}(d_0)$ is asymptotically a similar statistic for all nuisance parameters when the QTL is supposed to be located at d_0 and gives the asymptotic distribution of the statistic under an alternative hypothesis.

Proposition 1
Under the null hypothesis, *i.e.*, the QTL is located at d_0, we obtain

$$\mathbf{Z}(d_0) \overset{\lim}{\rightarrow} N(\mathbf{0},\mathbf{V}) \ ,$$

where $\mathbf{V}$ is a $(J\text{-}1)\times(J\text{-}1)$ symmetric matrix which depends only on the length of the chromosome and on the positions of the markers. Under an alternative hypothesis, *i.e.*, the QTL is located at d, we obtain

$$\mathbf{Z}(d_0) \overset{\lim}{\rightarrow} N\left[\mathbf{X}(d,d_0)\frac{\delta}{\sigma},\mathbf{V}\right] \ ,$$

where $\mathbf{X}(d,d_0)$ depends only on the length of the chromosome, on the positions of the markers, and on d and d_0.

Using the asymptotic distribution of $\mathbf{Z}(d_0)$ a maximum likelihood ratio test statistic denoted by $T_{\mathbf{Z}}(d_0)$ can be constructed. In the local asymptotic framework, the asymptotic distribution of $T_{\mathbf{Z}}(d_0)$ under the null hypothesis does not depend on the nuisance parameters: it is the distribution of the supremum of a χ_1^2 process with a covariance function depending on d_0 and the positions of the markers on the linkage map. In practical applications it is very difficult to obtain an explicit expression for the threshold function $c_\alpha(d_0)$, but values of this function can be obtained using simulation. Threshold functions for maps with equally spaced markers are given by Mangin *et al.* (1994).

Results and discussion

Because the newly proposed confidence interval is constructed using asymptotic arguments in a local asymptotic framework, it is important to check its qualities in real situations. This was done using simulation. Table 2 gives the coverage probability, *i.e.*, the probability that the confidence interval contains the true position of the QTL. It appears that only small deviations from the nominal value (90%) are found.

Table 2. Coverage probability of $T_Z(d_0)$ (in %)

% Variance due to QTL	$n = 200$		$n = 50$
	20 cM	5 cM	20 cM
0.5%	89.3	89.1	89.4
2%	90.1	90.0	89.0
5%	89.9	89.5	88.5
15%	89.9	89.7	88.7
50%	90.0	89.5	88.4
90%	89.5	89.4	88.6

Empirical coverage probabilities for the confidence interval based on $T_Z(d_0)$ based on 10,000 replications. In each simulation, a backcross progeny of 200 or 50 individuals (n) is generated with one 100 cM chromosome, 6 or 21 markers (20 cM or 5 cM, respectively) equally spaced, $\mu = 0$, $\sigma^2 = 1$ and a QTL in the middle of the chromosome. The threshold is taken to ensure a 90% confidence interval

Table 3 gives average lengths of confidence intervals and corresponding coverage probabilities (obtained by simulation) which can be compared with the results Darvasi *et al.* (1993). Means and coverage probabilities are calculated for all replications or conditionally on the fact that the QTL is detected using the LOD score using a type I error of 5 or 1%, respectively.

Darvasi *et al.* (1993) obtained by simulation a 95% confidence interval with lengths equal to 81 cM and 54 cM for $n = 500$ and $n = 1000$, respectively. With regard to the average length of the confidence interval, the newly proposed confidence interval is smaller for $n = 500$ compared to Darvasi's confidence interval. However, for replications which gave a LOD score larger than the threshold for a type I error of 5% or 1%, the average length of the interval decreases considerably whereas the coverage probability remains acceptable.

Table 3. Average length of new confidence interval (in cM) and coverage probabilities

$n = 500$		$n = 1000$	
Length	Coverage	Length	Coverage
All simulations:			
76	96%	55	96%
Only simulations with QTL detected (Type I error 5%):			
60	94%	50	96%
Only simulations with QTL detected (Type I error 1%):			
46	91%	42	95%

Empirical mean of the new confidence interval length and its coverage probability over 1000 replications. In each simulation, a backcross progeny is generated with a 100 cM chromosome, 6 markers equally spaced, $\mu = 0$, $\sigma^2 = 1$, $a = 0.25$ and a QTL in the middle of the chromosome. The threshold is taken to ensure a 95% confidence interval

The simulations can also be used to estimate the proportion of replications that gave a length smaller than 81 cM or 54 cM for $n=500$ or $n=1000$, respectively. For $n=500$, 49% of replications gave a length smaller 81 cM, whereas this proportion increases to 81% and 98% when using only those replications where a QTL is detected with a type I error of 5% and 1%, respectively. For $n=1000$, 55% of replications gave a confidence interval smaller than 54 cM, while this percentage increased to 64% and 78% when using only those replications where a QTL was detected using a type I error of 5% and 1%, respectively.

Although the computations involved in the newly proposed confidence interval look more complicated than those used for obtaining a classic confidence interval according to Lander & Botstein (1989) or an empirical confidence interval according to Darvasi *et al.* (1993), our method should be preferred because it guarantees an approximately unbiased confidence interval.

References

Conneally, P.M., J.H. Edwards, K.K. Kidd, J.M. Lalouel, N.E. Morton *et al.*, 1985. Reports of the committee methods of linkage analysis and reporting. Cytogenet. Cell Genet. 40: 356-359.

Cox, D.R. & D.V. Hinkley, 1974. Theoretical statistics. Chapman and Hall, London.

Darvasi, A., A. Weinreb, V. Minke, J.I. Weller & M. Soller, 1993. Detecting marker QTL gene effect and map location using a saturated genetic map. Genetics 134: 943-951.

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

Mangin, B., B. Goffinet & A. Rebaï, 1994. Constructing confidence intervals for QTL location. Genetics: in press.

Rebaï, A., B. Goffinet & B. Mangin, 1993. Comparing power of different methods for QTL detection. Biometrics: in press.

Sax, K., 1923. The association of sizes differences with seed coat pattern and pigmentation in *Phaseolus vulgarus*. Genetics 8: 552-560.

Three marker scanning of chromosomes for QTL in neighbouring intervals

O. Martínez & R.N. Curnow, Facultad de Ciencias, Universidad Autónoma de San Luis Potosí, México & Department of Applied Statistics, University of Reading, U.K.

Key words
markers, quantitative trait loci

Abstract

With sufficient data, interval mapping based on maximum likelihood methods or on regression methods should give satisfactory estimates of the location and size of a single QTL when it is the only substantial QTL in the whole genome. The standard errors of the estimates will be overestimated if QTL are segregating elsewhere in the genome. If the QTL are not in directly neighbouring intervals of the interval being investigated, this overestimation can be reduced by including cofactors or covariates representing the genotypes at other marker loci.

The problem that remains when there maybe QTL in directly neighbouring intervals is discussed. A QTL in a neighbouring interval can seriously bias the estimates of the sizes and locations of the QTL in the interval being investigated. The flanking markers of the other intervals cannot be used as cofactors because one of them is also a marker for the interval being studied. We recommend that regions of the genome that appear to contain QTL should be studied in greater detail by applying regression to each of the possible sets of three consecutive markers or three out of the four consecutive markers in the region. The residual sum of squares, minimised by choice of estimates of the size of the effects of the potential QTL in the neighbouring flanked regions, plotted as a bivariate surface against the possible positions for the QTL will allow decisions to be made about the possible presence of QTL in neighbouring flanked regions. Then a global analysis can be applied.

Introduction

The use of molecular markers in segregating populations of plants permits the estimation of the number, positions and sizes of effects of polygenes affecting quantitative characters. The positions are now referred to as Quantitative Trait Loci (QTLs). The estimation methods are based on the information supplied by the molecular markers given that we know the parent of origin of each marker allele.

Regression mapping consists of regressing the phenotype of the individuals in a segregating population on the probability of one or more QTLs segregating at particular positions between the markers, given that we know the marker genotype of each individual (Haley & Knott 1992, Martínez & Curnow 1992, 1994). In the simplest case, a pair of successive molecular markers is used and a single QTL is assumed to be segregating between them. Since we do not know the position of the QTL the estimation procedure consists in fitting the model by least squares for a grid of values of putative positions for the QTL. The resulting value of the residual sum of squares, RSS(t), is then graphed against all possible positions t between each pair of successive markers and for each chromosome. A clear minimum of the RSS(t) will indicate the presence of a QTL. The significance of a minimum can be tested by the usual F ratio of the regression analysis of variance. This procedure gives results remarkably similar to those of the maximum likelihood procedure of Lander & Botstein (1989), the graph of the RSS(t) being approximately inverse to the LOD score graph. The differences between the two methods are due to the regression method approximating the distribution of the phenotype within marker genotype groups as normal when in fact it is a mixture of normal distributions.

More than one QTL segregating

When there is only one QTL segregating in the whole genome and there are enough observations, either maximum likelihood or least squares methods give satisfactory estimates of the location and size of effect of the QTL. However, in general there will be an unknown number of QTLs scattered at various locations along the chromosomes. Both methods then present two problems: 1) When studying a particular marker interval, the estimated error variance will include genetic variance generated by QTLs segregating elsewhere in the genome, affecting the power of the test and overestimating the standard errors of the estimates; 2) If two or more QTLs are segregating in the same chromosome we may detect spurious QTLs (*ghost QTLs*) as well as incur severe biases in the estimated positions and sizes of effects of the real QTLs (Haley & Knott 1992, Jansen

1993, Martínez & Curnow 1992).

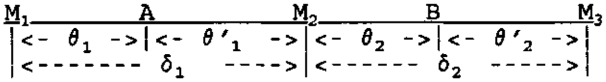
A method to solve these problems was proposed by Jansen and Stam (Jansen 1993, Jansen & Stam 1994). This procedure combines multiple linear regression with interval mapping and consists in fitting one QTL at a time in a given interval and simultaneously using some of the other markers as covariables to absorb the effects of other QTLs that could be segregating. As a first step in this procedure a subset of markers is selected as covariates by backward elimination. It is hoped that at least one marker selected by this procedure will be close to each of the segregating QTLs. In the second step of the procedure a search for QTLs is performed by means of interval mapping. This search proceeds interval by interval using subsets of the selected markers as covariates to absorb effects of QTLs that are not in the interval currently being searched. Akaike's Information Criterion is used to compare models with different numbers of degrees of freedom, and a difference of 2 in the values of this statistic for the two models is taken as significant. In this procedure, when fitting a QTL in a particular interval other previously selected markers in the same chromosome may be redundant. If all markers in the current chromosome are redundant when a QTL is fitted in a particular interval, this indicates that this QTL is the only QTL in the chromosome. Otherwise, if some markers in the current chromosome are still significant, even when a QTL has been fitted in the model, then the presence of multiple QTLs in that chromosome is indicated. Regression mapping could be used instead of interval mapping in this procedure.

The procedure almost eliminates the first problem, the overestimation of the error variance, but as we will see it still gives biased estimates of the positions and sizes of effects of QTLs when two of them are segregating in neighbouring regions of the same chromosome.

Here we present a method that enables discrimination between one and two QTLs in the same chromosome and, in the case of two QTLs in neighbouring intervals, correctly estimates the positions and sizes of effects of these genes. We base our approach on a regression model using three markers and searching for two QTLs at the same time. To study the average behaviour of this model we will use the expected residual sum of squares under different situations.

Regression mapping with three markers

Assume the situation



where M_1, M_2, M_3 are markers at distances apart δ_1 and δ_2, A and B are QTLs, and θ_1, θ'_1, θ_2 and θ'_2 are the recombination probabilities between the various loci. Assume also that a backcross population B_1 is available; $B_1 = F_1 \times P_1$, where the genotypes of the individuals in the F_1 population are $F_1 = M_1AM_2BM_3 / m_1am_2bm_3$, and in the P_1 population $P_1 = m_1am_2bm_3 / m_1am_2bm_3$. In this B_1 population we can distinguish eight different marker classes, $i = 1, 2, \dots 8$. For each marker class we have four possible QTL genotypes, that can again be labelled by the gamete from the F_1 population as AB, Ab, aB and ab. All the gametes from the P_1 parent will have ab at the QTLs. Now consider the expectation of the value of the character in the j-th individual that belongs to the i-th marker class for $j = 1, 2, \dots n_i$; $i = 1, 2, \dots 8$, say

$$\begin{aligned} E[Y_{ij}] &= \beta_0 + \beta_1 P[AB \mid i] + \beta_2 P[Ab \mid i] + \beta_3 P[aB \mid i] \\ &= \beta_0 + \beta_1 \Delta_i + \beta_2 \Lambda_i + \beta_3 \Psi_i , \end{aligned} \tag{1}$$

where β_1, β_2 and β_3 are the effects of the QTL genotypes AB/ab, Ab/ab and aB/ab respectively, relative to a baseline the β_0, the general mean of the population plus the value of the QTL genotype ab/ab; Δ_i, Λ_i and Ψ_i are the conditional probabilities of inheriting from the F_1 parent the QTL gametes AB, Ab and aB respectively given the marker class. Then we consider the model

$$Y_{ij} = \beta_0 + \beta_1 \Delta_i + \beta_2 \Lambda_i + \beta_3 \Psi_i + \varepsilon_{ij}^{*} , \tag{2}$$

where ε_{ij}^{*} is an error term. The probabilities Δ_i, Λ_i and Ψ_i are known functions of the unknown values of the parameters θ_1 and θ_2.

Because we do not know the values of θ_1 and θ_2 we consider, instead of (2), the model

$$Y_{ij} = \beta_0 + \beta_1 \Delta_i(\mathbf{t}) + \beta_2 \Lambda_i(\mathbf{t}) + \beta_3 \Psi_i(\mathbf{t}) + \varepsilon_{ij} , \tag{3}$$

where $\Delta_i(\mathbf{t})$, $\Lambda_i(\mathbf{t})$ and $\Psi_i(\mathbf{t})$ are the corresponding functions Δ_i, Λ_i and Ψ_i now evaluated at $\mathbf{t} = (t_1, t_2)$, where $0 \leq t_1 \leq \delta_1$, $0 \leq t_2 \leq \delta_2$; that is, t_1 and t_2 are putative values for θ_1 and θ_2, respectively. If we estimate the parameters β_0, β_1, β_2 and β_3 in (3) by least squares we obtain, for a given value of $\mathbf{t}$, the predicted value of y_{ij}, the same for all j,

$$\hat{y}_{ij}(\mathbf{t}) = \hat{\beta}_0(\mathbf{t}) + \hat{\beta}_1(\mathbf{t})\Delta_i(\mathbf{t}) + \hat{\beta}_2(\mathbf{t})\Lambda_i(\mathbf{t}) + \hat{\beta}_3(\mathbf{t})\Psi_i(\mathbf{t}) \;, \tag{4}$$

A measure of the fit of the model, for each value of $\mathbf{t}$, is given by the residual sum of squares,

$$\mathrm{RSS}(\mathbf{t}) = \sum_{i=1}^{8} \sum_{j=1}^{j=n_i} (y_{ij} - \hat{y}_{ij}(\mathbf{t}))^2 \;, \tag{5}$$

The estimator of $\boldsymbol{\theta} = (\theta_1,\ \theta_2)$ is the value of $\mathbf{t}$ on the bivariate surface $\{\, 0 \le t_1 \le \delta_1,\ 0 \le t_2 \le \delta_2 \,\}$ that minimizes the residual sum of squares given by (5), say $\mathbf{t} = \mathbf{t}_m$.

We will use the expected value of RSS($\mathbf{t}$),

$$\mathrm{E}[\mathrm{RSS}(\mathbf{t})] = \mathrm{E}[\sum_{i=1}^{8} \sum_{j=1}^{N_i} (Y_{ij} - \hat{Y}_{ij}(\mathbf{t}))^2] \;. \tag{6}$$

Neglecting terms that result from the variance of the marker class means, (6) can be approximated by

$$\mathrm{E}[\mathrm{RSS}(\mathbf{t})] \approx n \left\{ \sum_{i=1}^{8} p_i V[Y_i] + \sum_{i=1}^{8} p_i \left(\mathrm{E}[Y_{ij}] - \mathrm{E}[\hat{Y}_{ij}(\mathbf{t})]\right)^2 \right\}$$

$$= n \{\, V_w + h(\mathbf{t}) \,\}, \tag{7}$$

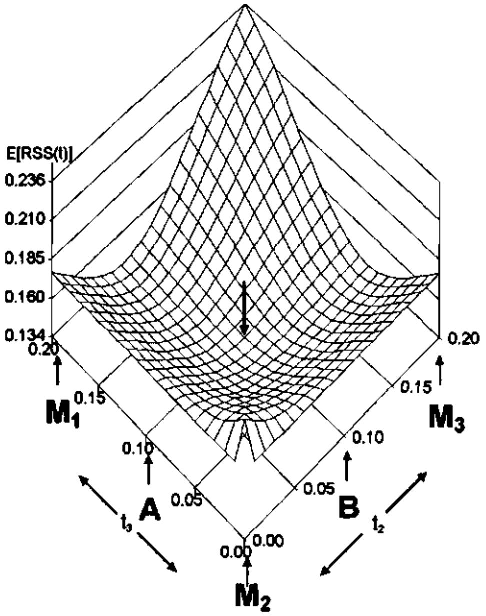
where the constant n is the fixed sample size, and can be ignored in the minimization; the values p_i are the probabilities of each one of the marker classes, $i = 1, 2, \ldots 8$; the term V_w is the weighted variance of the character within marker classes and

$$h(\mathbf{t}) = \sum_{i=1}^{8} p_i \left(\mathrm{E}[Y_{ij}] - \mathrm{E}[\hat{Y}_{ij}(\mathbf{t})]\right)^2 \;. \tag{8}$$

$h(\mathbf{t})$ depends on $\mathbf{t}$ as well as on the markers to be used in the model.

We will study E[RSS($\mathbf{t}$)] given by (7) under two contrasting situations: a) there are two QTLs in neighbouring intervals, say: $M_1AM_2BM_3$; and b) there is only one QTL in one of the intervals, say: $M_1AM_2M_3$. Figure 1 presents the graph of E[RSS($\mathbf{t}$)], strictly $V_w + h(\mathbf{t})$ and ignoring the constant error term in V_w, for the three marker regression

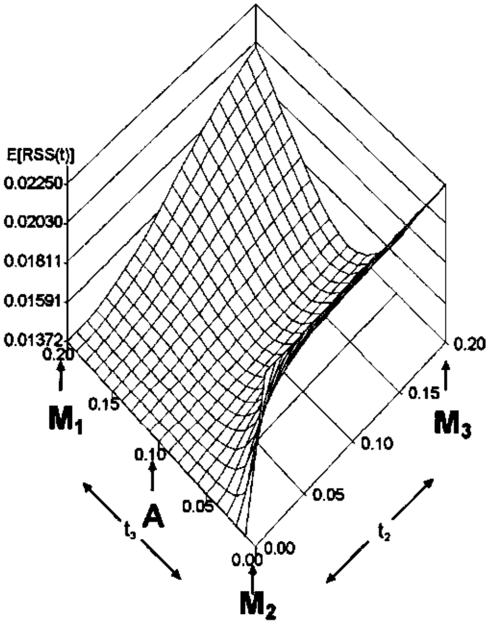
Figure 1. Expected residual sum of squares for the three markers regression mapping model assuming two QTLs when two QTLs, A and B, are present. Effects relative to the effect of the ab gamete; Ab ½, aB ½, AB 2



model under situation a), two QTLs, A and B present, one in the middle of each interval with equal sizes and signs of effects and some epistasis. Figure 2 presents the graph of E[RSS(**t**)] for the three marker regression model under situation b), only one QTL, A, is present and it is in the first interval. In Figures 1 and 2, the variable $t_3 = 0.2 - t_1$ is used in place of t_1 to aid interpretation. In Figure 1 the E[RSS(**t**)] surface presents a global minimum, exactly at the positions of the two QTLs. At this point the expected values of the estimates of the sizes of the effects coincide with the true values. Figure 2 presents two lines of multiple minima, the first along the line $t_3 = 0.1$, $0 < t_2 < 0.2$, and the second along the line $t_2 = 0$, $0 < t_3 < 0.2$. This indicates the presence of only one QTL in the first interval (M$_1$-M$_2$), since, when this QTL is assumed, the assumption of another QTL in the second interval (M$_2$-M$_3$) does not improve the fit of the model. Along these lines the expected value of the size of the effect of the QTL is equal to its true value. In summary, discrimination between two or one QTLs is possible by observing the presence of a single minima in the E[RSS(**t**)] surface or multiple minima along two straight lines. In the next section we explain how this can be used when solving multiple minima in the graph of the RSS(**t**), or multiple maxima in the LOD(**t**) score graph when using interval mapping with pairs of markers.

The generalization of Regression Mapping to include models with more than two

Figure 2. Expected residual sum of squares for the three markers regression mapping model assuming two QTLs when in fact only one, A, is present. The effect of the A gamete relative to the a gamete, A ½



QTLs at the same time is straightforward. However, this generalization is not very useful in practice because of the very large sample size required and the difficulties of interpreting the shape of a multidimensional RSS(**t**) surface.

Solving multiple minima

As mentioned before, when there is only one QTL in a given interval the most likely event is that a single minimum appears near the true position of the QTL in the interval that contains it. However, even in this simple case spurious minima in neighbouring intervals do occur. In general the graph of RSS(**t**) for a given chromosome can present more than one significant minima and then we need a procedure to decide between the hypothesis of one or two QTLs. The procedure that we propose is to use the results of the analyses performed with all the possible sets of three markers that are contained in the region where the multiple minima are present. With sufficient data these analyses will be logically consistent with only one of two hypotheses: one or two QTLs. If there are three or more QTLs in the region this procedure will be inconclusive and only a regression mapping model of higher dimension can be decisive.

Denote by M_l and M_k the extreme markers that contain the significant minima, where successive markers are numerated by the subindexes $1, 2, \ldots k$. We can perform $k-2$ different three marker analyses using as flanking markers M_1 and M_k, and as the middle marker M_2, M_3, ... M_{k-1}; ignoring in each case the other markers. If there is only one QTL segregating in one of the $k-1$ intervals, all the analyses will give a RSS(**t**) graph with shape as in Figure 2, with multiple minima along two lines; one along the interval where the QTL is located and the other perpendicular to this line, indicating the estimated position of the QTL. If there are two QTLs, the graphs of the RSS(**t**) where the QTLs are in different intervals will show a single minimum, estimating the position of both QTLs. When two QTLs are located in the same interval, they behave for estimation purposes as a single QTL with position and size of effect determined by the real positions and sizes of effects of the two QTLs. Then the graphs where the two QTLs are located in the same interval will have the same shape as Figure 2.

As an example assume that there is more than one minimum present in the interval flanked by markers M_1 to M_5, and assume that the real situation is $M_1M_2M_3QM_4M_5$; that is, there is a single QTL segregating between M_3 and M_4. In this case, all the analyses using M_1-M_2-M_5, M_1-M_3-M_5 and M_1-M_4-M_5 will show two line minima, consistent with the true position of the QTL between M_3-M_4. Even when the estimated position in each analysis could be different by random variation, the shape of the RSS(**t**) surface will be consistent with the hypothesis of only one QTL. In this case we will use the estimates of position and size of effect given by the analysis using only one pair of markers (M_3 and M_4), because these estimates will be more precise.

Now, assume that there is more than one single minimum in the interval flanked by markers M_1 to M_5, but now the real situation is $M_1Q_1M_2M_3Q_2M_4M_5$; that is, there are two QTLs, Q_1 in the first interval and Q_2 in the third. The analysis using M_1-M_2-M_5 will show a single minimum, indicating the presence of two QTLs. The analysis using M_1-M_3-M_5 will also show a single minimum, consistent with the hypothesis of two QTLs, but the analysis using M_1-M_4-M_5 will show two line minima, indicating evidence of only one QTL. This is because the two QTLs Q_1 and Q_2 are included in the same interval, formed by M_1-M_4 and are therefore not separable by a model that allows only one QTL in that interval. In this case the evidence points to two QTLs in the correct intervals: M_1-M_2 and M_3-M_4. The best possible estimates for this situation are obtained from the model including two QTLs one in each of these intervals. This model is easily constructed, regressing the probabilities of each QTL genotype, given the marker genotypes of markers M_1, M_2, M_3 and M_4, and performing a bivariate search in the surface generated when taking into account the two putative positions of the QTLs.

In all cases extra analyses can be performed to reinforce the conclusion reached; for example in the situation $M_1M_2M_3QM_4M_5$ an analysis using markers $M_1M_2M_3$ will indicate, correctly, that the best fit of the model is assuming only one QTL completely linked with the marker M_3. This is because the QTL is beyond the region searched by the model. In the situation $M_1Q_1M_2M_3Q_2M_4M_5$ it is possible to perform analyses, using for example M_1-M_2-M_3 and M_2-M_3-M_4. If the signs of the effects of the QTLs Q_1 and Q_2 are the same, then a *ghost* QTL will be present between M_2 and M_3, and both analyses will provide evidence of two QTLs, but the estimated position of the ghost QTL between M_2 and M_3 will move when estimated with different markers, indicating a spurious effect.

The algorithm we have proposed to estimate multiple QTLs solves the problems of multiple minima and provides consistent estimates of the number, positions and sizes of effects of multiple QTLs. The procedure consist of three steps: a) Locating significant minima using pairs of markers, b) Solving multiple minima by three markers regression mapping and c) Fitting a global model including all significant QTLs and their effects. We have presented the step b) in detail. The third step, c) consists in fitting a model that includes all the QTLs found in the different chromosomes, after solving the problems given by multiple minima. If the global model is correct, that is if it includes all the real QTLs, the estimated error variance will be correct, and the standard error of the estimators of the sizes of effect will not be overestimated. Also in this global model the sizes of any epistatic interactions there may be between QTLs can be estimated and tested for significance.

Discussion

We have mentioned that the procedure of Jansen (1993) and Jansen & Stam (1994) will give biased estimates of the positions and sizes of effects when there are two QTLs in neighbouring flanked intervals. This is easy to see if we note that using markers as cofactors is equivalent to fitting a model with a QTL at the marker position. To deal with the situation $M_1Q_1M_2Q_2M_3$ Jansen & Stam will, in turn, perform interval mapping between M_1 and M_2 using M_3 as cofactor, and then perform interval mapping between M_2 and M_3 using M_1 as cofactor. Figure 1 shows that the estimates obtained by Jansen & Stam will be biased. Performing interval mapping or regression mapping between M_1 and M_2 using M_3 as cofactor corresponds to moving between M_1-M_2 with a QTL fixed at M_3; i.e. we will be observing the values of E[RSS(t)] along the line $\{0 < t_3 < 0.2, t_2 = 0.2\}$. Note that the minimum of the E[RSS(t)] along this line is not at the true position of the QTL A (0.1), but shifted towards M_2, at about $t_3 = 0.07$. The expected values of the estimates of the gametic effects relative to the effect of the ab gamete will be AB 1.98, Ab 1.22 and aB 0.21 compared with their true values of 2, 0.5 and 0.5. With an additive system and model, the expected values of the estimates of the effects of the A and B gametes relative to the a and b gametes will be 0.74 compared with a true value of 0.5 and the estimated positions of the loci will again be at about $t_3 = 0.07$ compared with the true value of 0.10. The biases in the estimated positions occur because recombinations between the QTL and the marker used as a cofactor are not included in the univariate search. There would be no bias if the QTL that is not included in the search is completely linked with the marker used as cofactor, and the bias will be a maximum when the QTL is far distant from this marker.

In conclusion, when there maybe two QTLs in neighbouring intervals the correct way to estimate their positions and sizes of effect is to perform a bivariate search in both intervals at the same time. The application of the methods we have described to the analysis of data from a set of double haploid lines obtained from the F_1 generation of a cross between two lines of wheat (*Triticum aestivum*) will be described in a further publication.

References

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

Jansen, R.C., 1993. Interval mapping of multiple quantitative trait loci. Genetics 135: 205-211

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

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

Martinez, O. & R.N. Curnow, 1992. Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. Theor. Appl. Genet. 85: 480-488

Martinez, O. & R.N. Curnow, 1994. Missing markers when estimating quantitative trait loci using regression mapping. Heredity 73: 198-206

Heteroscedasticity in multilocation trials: Implications for stability analysis and the ANOVA *F*-test

Hans-Peter Piepho, Faculty of Agriculture, University of Kassel, 37213 Witzenhausen, Germany

Introduction

The presence of genotype-environment interaction is common in multilocation yield trial data. Genotypes that show large interactions with environments are often judged to be unstable. Estimating variance components attributable to the interactions of single cultivars is therefore a useful means of assessing phenotypic stability. The stability variance suggested by Shukla (1972a) is one such measure. It assesses the interaction plus the mean error variance of a genotype. Error variances are assumed to be homogeneous, so that differences in stability variances are merely due to differences in interaction variances. A small stability variance is then indicative of a high stability. Several statistical tests for the global hypothesis of no stability differences among genotypes and for multiple comparisons are available.

The stability variance is an appropriate measure only if the error variances are homogeneous across genotypes and the number of replicates is the same in each environment (Shukla 1972). Otherwise one does not know, whether differences in stability variance are due to heteroscedastic interactions or due to heteroscedastic errors. A distinction between these two sources of heteroscedasticity is necessary, if stability is to be defined in terms of interaction. An alternative estimation procedure has been suggested for the case that Shukla's assumptions are violated (Piepho 1994a). It yields separate estimates for the interaction variance and the error variance of a genotype.

Stability differences as well as heterogeneity of error variances have an influence on the comparison of yield means. Data from yield trials, conducted in different environments, are frequently analysed by a combined analysis of variance. One of the assumptions underlying such analyses is that error variances be homogeneous across treatments and environments. This assumption may not always be valid (Cochran & Cox 1957). If a mixed model with fixed genotypes and random environments is assumed, genotype-environment interaction is a random effect. For the ordinary ANOVA to be valid in this case, it is then also required that stability variances be homogeneous across genotypes. If there are stability differences among genotypes, this condition is violated. This contribution suggests Box's correction of the ANOVA degrees of freedom for the case of heterogeneous interaction and error variances.

Model

The analysis of phenotypic stability is often based on a two-way mixed model of the form (Shukla 1972a)

$$x_{ijm} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijm} ,$$

where x_{ijm} is the observation of the phenotypic value of genotype i $(= 1, 2, ..., K)$ in replication m $(= 1, 2, ... , R)$ of environment j $(= 1, 2, ..., N)$, μ is the overall mean, α_i is the fixed effect of genotype i, β_j is the random effect of environment j, $(\alpha\beta)_{ij}$ is the random interaction effect of genotype i and environment j, and e_{ijm} is the experimental error associated with x_{ijm}. It is assumed that all random effects are independently distributed with zero mean. Stability statistics are usually based on means, for which the model reads

$$x_{ij} = \mu + \alpha_i + \beta_j + v_{ij} , \qquad (1)$$

where $x_{ij} = \Sigma_m\, x_{ijm}/R$ and $v_{ij} = (\alpha\beta)_{ij} + \Sigma_m\, e_{ijm}/R$. Note that with the means model it is no longer possible to distinguish between genotype-environment interaction and experimental error.

Strictly speaking, the model used by Shukla (1972a) is appropriate only for data from yield trials laid out as a completely randomized (CR) design. A much more common design prevailing in most multilocation trials is the randomized complete block (RCB) design, for which the linear model reads

$$x_{ijm} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \tau_{jm} + e_{ijm} , \qquad (2)$$

where τ_{jm} is the effect of block m in environment j. The corresponding means model is given by

$$x_{ij} = \mu + \alpha + \beta_j^* + v_{ij} ,$$

where $\beta_j^* = \beta_j + \Sigma_m\, \tau_{jm}/R$. The only difference to eqn (1) lies in the addition of the mean of blocks to the environmental effect. β_j^* may simply be regarded as a modified environmental effect.

A genotype is stable, if the "stability variance" σ_i^2, *i.e.*, the variance of the effects v_{ij}, is small. The term stability variance for $\sigma_i^2 = \text{var}(v_{ij})$ was coined by Shukla (1972a). For other concepts of stability see, *e.g.*, Lin *et al.* (1986). If we assume that all genotypes have a common error variance $\delta^2 = \text{var}(e_{ijm})$, stability differences result merely from differences of a genotype's interaction variance $\theta_i^2 = \text{var}((\alpha\beta)_{ij})$. The stability variance can be expressed as $\sigma_i^2 = \theta_i^2 + \delta^2/R$. Maximum stability occurs when $\theta_i^2 = 0$ and hence $\sigma_i^2 = \delta^2/R$, *i.e.*, when the variability of v_{ij} effects is minimal. The statistical design has an influence on the magnitude of δ^2 and hence on the magnitude of σ_i^2 (δ^2 and σ_i^2 will tend to be smaller for an RCB design than for a CR design), but it does not influence contrasts among the stability variances of any two genotypes. Clearly, the difference $\sigma_r^2 - \sigma_s^2$ equals $\theta_r^2 - \theta_s^2$ for any r, $s = 1, 2, \ldots, K$, which is independent of δ^2.

Tests of stability differences with homoscedasticity of errors

Several tests for equality of stability variances are available (Table 1). Simulations by Piepho (1992) have shown that the parametric global (H_0: $\sigma_i^2 = \sigma^2$) and pairwise (H_0: $\sigma_i^2 = \sigma_{i'}^2$) tests are rather sensitive to departures from normality. It is therefore suggested to use robust or rank procedures.

Estimating stability under heteroscedasticity

If the error variance is not the same for each genotype, *i.e.*, $\text{var}(e_{ijm}) = \delta_i^2$, estimating stability by the above procedures is not adequate. It may then be more appropriate to estimate $\theta_i^2 = \text{var}((\alpha\beta)_{ij})$, rather than $\sigma_i^2 = \text{var}(v_{ij})$, for each genotype. If in each

Table 1. Tests for equality of stability variances

H_0: $\sigma_i^2 = \sigma^2$ for every i	H_0: $\sigma_i^2 = \sigma_{i'}^2$ for $i \neq i'$
I. Parametric tests	
Anscombe (1981)	Johnson (1962)
Shukla (1982)	Maloney & Rastogi (1970)
Brindley & Bradley (1985)	Shukla (1972b)
Mudholkar & Sarkar (1992)	
II. Robust tests	
Levene (1960)	t-test based on $\mid v_{ij} \mid$
Piepho (1994b)	
III. Rank tests	
Nassar & Hühn (1987)/ Hühn & Nassar (1989)	?
Piepho (1994c)	

environment the yield trial was laid out as a RCB design with R replications, estimates of δ_i^2 and θ_i^2 are given by

$$\hat{\delta}_i^2 = \frac{K(K-1)U_i - \sum_i U_i}{N(R-1)(K-1)(K-2)} ,$$

$$\text{where} \quad U_i = \sum_j \sum_m (x_{ijm} - \bar{x}_{ij.} - \bar{x}_{.jm} + \bar{x}_{.j.})^2 \text{ , and } \quad \hat{\theta}_i^2 = S_{vi}^2 - \frac{\hat{\delta}_i^2}{R} ,$$

$$\text{where} \quad S_{vi}^2 = \frac{K(K-1)Z_i - \sum_i Z_i}{(N-1)(K-1)(K-2)} \text{ , with } \quad Z_i = \sum_j (\bar{x}_{ij.} - \bar{x}_{i..} - \bar{x}_{.j.} + \bar{x}_{...})^2 .$$

Implications of heteroscedasticity for *F*-tests

The ANOVA table for the analysis of a yield experiment laid out as a RCB design is shown in Table 2. The two F-tests of particular interest are those for significant genotypic effects ($F_1 = MS_3/MS_4$) and for significant genotype-environment interaction ($F_2 = MS_4/MS_5$). Under homoscedasticity, F_1 is distributed as an F distribution with $K-1$ and $(K-1)(N-1)$ degrees of freedom, whereas F_2 is distributed as an F distribution with $(K-1)(N-1)$ and $N(K-1)(R-1)$ degrees of freedom. In the presence of heteroscedasticity, however, these tests are not valid.

Considering the four types of heteroscedasticity outlined in Table 3, the following consequences for these two F-tests are possible:

Table 2. ANOVA table for yield trials of K genotypes (fixed) laid out as RCB designs with R replicates and conducted in N environments (random)

Source	df	MS	E(MS)
Environments	$N-1$	MS_1	$\delta^2 + R\theta^2 + K\sigma_\tau^2 + KR\sigma_\beta^2$
Blocks (Env.)	$N(R-1)$	MS_2	$\delta^2 + K\sigma_\tau^2$
Genotypes	$K-1$	MS_3	$\delta^2 + R\theta^2 + NR\Phi(\alpha)$
Genotype × Env.	$(K-1)(N-1)$	MS_4	$\delta^2 + R\theta^2$
Error	$N(K-1)(R-1)$	MS_5	δ^2

δ^2 = error variance; θ^2 = interaction variance; σ_τ^2 = block variance; σ_β^2 = environmental variance; $\Phi(\alpha) = \Sigma_i\, \alpha_i^2/(K-1)$

Table 3. Types of heteroscedasticity

		Stability variance σ_i^2	
		homogeneous	heterogeneous
Error variance δ_i^2	homogeneous	a	c
	heterogeneous	b	d

(a) Ordinary ANOVA valid,

(b) F-test for interactions not valid,

(c) F-test for genotypic effects not valid,

(d) F-tests for genotypic effects and for interactions not valid.

Only in case (a) the statistical analysis may proceed as usual. In all other cases, assumptions of the usual ANOVA are violated, and we have to adjust the degrees of freedom by Box's method (or use some other robust or nonparametric procedure). Case (b) implies that heterogeneity among the variances θ_i^2 and δ_i^2 is such that $\theta_i^2 + \delta_i^2/R = \sigma^2$ for every i, which is quite unlikely. It is therefore suggested to regard σ_i^2 as heterogeneous whenever heterogeneity is detected in the error variances δ_i^2, regardless of the outcome of a test of homoscedasticity for σ_i^2. The following approach is proposed here:

(1) Test for homogeneity of error variances δ_i^2.

(2) If homogeneity of δ_i^2 is rejected, adjust the degrees of freedom of the F-tests for interaction and genotypic effects (see below). Otherwise test for homogeneity of σ_i^2.

(3) If homogeneity of σ_i^2 is rejected, adjust degrees of freedom of F-tests for genotypic effects (see below). Otherwise conduct ordinary ANOVA.

Homogeneity of genotypic error variances δ_i^2 may be tested by subjecting the block experiment of each environment to a test of homoscedasticity in a two-way layout. To circumvent the problem of multiple testing, the test statistics in each environment may be combined into one single statistic. Two tests seem appropriate for this procedure, namely those introduced by Shukla (1982) and by Brindley & Bradley (1985), which yield independent chi-squared statistics for each environment.

Results by Box (1954) and Geisser & Greenhouse (1958) show that in case of heteroscedasticity F_1 is distributed as an F distribution with $\epsilon(K-1)$ and $\epsilon(K-1)(N-1)$ degrees of freedom, whereas F_2 is distributed as an F distribution with $\epsilon'(K-1)(N-1)$ and $\epsilon' N(K-1)(R-1)$ degrees of freedom, where ϵ and ϵ' are correction factors for the degrees of freedom (ϵ, $\epsilon' \leq 1.0$) given by

$$\epsilon = \frac{(K-1)\sigma^4}{(K-2)K^{-1}\sum_i \sigma_i^4 + \sigma^4}, \text{ and } \epsilon' = \frac{(K-1)\delta^4}{(K-2)K^{-1}\sum_i \delta_i^4 + \delta^4},$$

where $\sigma^4 = K^{-2}(\Sigma_i\ \sigma_i^2)^2$ and $\delta^4 = K^{-2}(\Sigma_i\ \delta_i^2)^2$. Since ϵ and ϵ' are bounded downwards by a minimum value of $(K-1)^{-1}$, we may conduct the usual F-tests with degrees of freedom reduced by a factor of $(K-1)^{-1}$ (Cochran & Cox 1957: 551). This procedure is known as Box's conservative test. A gain in power is possible by using appropriate estimators of ϵ and ϵ'. Following a suggestion by Greenhouse & Geisser (1959) we may proceed as shown in Table 4 (see also Milliken & Johnson 1984), if a preliminary test reveals heteroscedasticity relevant for either the test based on F_1 or the test based on F_2.

If no significance is attained based on the usual degrees of freedom (step 1), the analysis ends, since the test based on adjusted degrees of freedom will also be nonsignificant. Otherwise proceed to step 2. If Box's conservative test (step 2) is significant, the analysis may be terminated, since the test based on adjusted degrees of freedom will also be significant. Otherwise conduct the F-test using Box's correction (step 3). The correction factors ϵ and ϵ' have to be estimated from the data. Several estimates are discussed by Piepho (1994d).

Similar procedures may be used to adjust the degrees of freedom in case the error variances vary among environments. The test based on F_2 with degrees of freedom adjusted for this case was shown to compare favourably well to the approximate procedure suggested by Cochran & Cox (1957), which tends to be conservative (Piepho 1994d).

References

Anscombe, F.J., 1981. Computing in Statistical Sciences through APL. Springer Series in Statistics. Springer-Verlag, New York.

Box, G.E.P., 1954. Some theorems on quadratic forms in the study of analysis of variance problems. I. Effect

Table 4. Numerator and denominator degrees of freedom for F-tests following the three step procedure of Greenhouse & Geisser

	F_1		F_2	
	numerator	denominator	numerator	denominator
Step 1: Test with usual df	$(K-1)$	$(K-1)(N-1)$	$(K-1)(N-1)$	$N(K-1)(R-1)$
Step 2: Box's conservative test	1	$(N-1)$	$(N-1)$	$N(R-1)$
Step 3: Test using Box's ϵ	$\epsilon(K-1)$	$\epsilon(K-1)(N-1)$	$\epsilon'(K-1)(N-1)$	$\epsilon' N(K-1)(R-1)$

of inequality of variance in the one-way classification. II. Effect of inequality of variance and correlation between errors in the two-way classification. Annals of Mathematical Statistics 25: 290-302, 484-498.

Brindley, R.D. & R.A. Bradley, 1985. Some new results on Grubbs' estimates. Journal of the American Statistical Association 80: 711-714.

Cochran, W.G. & G.M. Cox, 1957. Experimental design. Wiley, New York.

Geisser, S. & S.W. Greenhouse, 1958. An extension of Box's results on the use of the F-distribution in multivariate analysis. Annals of Mathematical Statistics 29: 885-891.

Greenhouse, S.W. & S. Geisser, 1959. On methods in the analysis of profile data. Psychometrika 24: 95-112.

Hühn, M. & R. Nassar, 1989. On tests of significance for nonparametric measures of phenotypic stability. Biometrics 45: 997-1000.

Johnson, N.L., 1962. Some notes on the investigation of heterogeneity in interactions. Trabajos de estadistica XIII: 183-199.

Levene, H., 1960. Robust tests for equality of variances. In: I. Olkin, S.G. Ghurye, W. Hoeffding, W.G. Madow & H.B. Mann (Eds.). Contributions to probability and statistics. Essays in honor of Harold Hotelling. Stanford University Press, Stanford, Carolina, 278-292.

Lin, C.S., M.R. Binns & L.P. Levkovitch, 1986. Stability analysis: where do we stand? Crop Science 26: 894-900.

Mudholkar, G.S. & I.C. Sarkar, 1992. Testing homoscedasticity in a two-way table. Biometrics 48: 883-888.

Maloney, C.J. & S.C. Rastogi, 1970. Significance tests for Grubbs' estimators. Biometrics 26: 671-676.

Milliken, G.A. & D.E. Johnson, 1984. Analysis of messy data. Vol I. Designed experiments. New York: Van Nostrand Reinhold.

Nassar, R. & M. Hühn, 1987. Studies on estimation of phenotypic stability: Tests of significance for nonparametric measures of phenotypic stability. Biometrics 43: 45-53.

Piepho, H.P., 1992. Vergleichende Untersuchungen der statistischen Eigenschaften verschiedener Stabilitätsmaβe mit Anwendungen auf Hafer, Winterraps, Ackerbohnen sowie Futter- und Zuckerrüben. Dissertation, Kiel, 163 p.

Piepho, H.P., 1994a. Application of a generalized Grubbs model in the analysis of genotype-environment interaction. To appear in Heredity.

Piepho, H.P., 1994b. A robust test for homoscedasticity in a two-way layout. To appear in Biometrical Journal.

Piepho, H.P., 1994c. A distribution-free test for homoscedasticity in a two-way layout. To appear in Journal of Statistical Computation and Simulation.

Piepho, H.P., 1994d. Detecting, interpreting and handling heteroscedasticity in yield trial data. Submitted.

Shukla, G.K., 1972a. Some statistical aspects of partitioning genotype-environmental components of variability. Heredity, 29: 237-245.

Shukla, G.K., 1972b. An invariant test for homogeneity of variances in a two-way classification. Biometrics 28: 1063-1072.

Shukla, G.K., 1982. Testing the heterogeneity of variances in a two-way classification. Biometrika 69: 411-416.

Detecting QTLs with diallel schemes

A. Rebaï, B. Goffinet, B. Mangin & D. Perret[*]*, INRA, Centre de Toulouse, Station de Biometrie et d'Intelligence Artificielle, Chemin Borde-Rouge, Auzeville B.P. 27, 31326 Castanet-Tolosan,* [*]*Rustica semences, Domaine de Sandreau, 31000 Toulouse, France*

Abstract

We describe a general interval mapping method for QTL detection using progenies derived from several connected F_2 populations issuing from diallel crosses among different lines. Linear model-based procedures are used for the test and estimation of putative QTL effects in such designs. Genetic interactions including epistasis are also investigated. The method is implemented and applied on simulated and experimental data from a 4×4 diallel of corn. Results show the consistency and the good power of the tests used.

Introduction

Many powerful methods, using information from pairs of neighbouring markers, have been proposed for the mapping of QTL. We have shown (Rebaï *et al.* 1994) that the interval mapping method of Lander & Botstein (1989) and the linear approach (Knapp *et al.* 1990, Haley & Knott 1992) have similar power for large population sizes. However, the last one provides models which are easier to generalize to complex experimental designs as those involving several connected populations.

In this paper we describe a widely applicable method for QTL detection using populations derived from crosses between different lines. The approach is demonstrated for the case of a diallel cross between 4 inbreds (L_1 to L_4) with no selfings nor reciprocals. Six F_2 populations were obtained and genotyped with respect to RFLP markers. F_3 families, obtained by selfing F_2 individuals, were scored for many characters and crossed to the two non parental lines as testers (F_3 from $L_1 \times L_2$ are crossed with L_3 and L_4). *TC* progenies, so obtained, are measured in different environments.

Models and Tests

The models and tests using one marker at a time were described in Rebaï & Goffinet (1993). These can be generalized to pairs of adjacent markers so as to get what is called the interval mapping approach originally proposed by Lander & Botstein (1989). At each position on the genome, flanked by codominant markers with known genotypes, we are able to write the expectations of marker classes means under the hypothesis that a QTL lies in that position. One just needs to calculate the conditional probabilities of the putative QTL genotypes as a function of recombination rates between QTL and markers (Knapp *et al.* 1990, Knott & Haley 1992). So, for each individual and every position on the genome one can write the expected phenotype assuming the presence of a QTL in that position and regress the observed phenotype on the QTL parameters (additivity, dominance) to estimate and to test them. Let us consider an F_2 population derived from the cross $L_i \times L_j$ with marker alleles indexed i and j, respectively. Consider two linked markers A and B and a putative QTL Q between them, then we have nine marker classes with expectations:

$$\begin{array}{lcllll}
\theta_1(A_iA_iB_iB_i) & = & \mu_{ij} + & 2(1-s)a_i + & 2sa_j + & 2s(1-s)d_{ij} \\
\theta_2(A_iA_iB_iB_j) & = & \mu_{ij} + & (2-s-t)a_i + & (s+t)a_j + & (s+t-2st)d_{ij} \\
\theta_3(A_iA_iB_jB_j) & = & \mu_{ij} + & 2(1-t)a_i + & 2ta_j + & 2t(1-t)d_{ij} \\
\theta_4(A_iA_jB_iB_i) & = & \mu_{ij} + & (1-s+t)a_i + & (1+s-t)a_j + & (1-s-t+2st)d_{ij} \\
\theta_5(A_iA_jB_iB_j) & = & \mu_{ij} + & a_i + & a_j + & [1-s(1-s)-t(1-t)]d_{ij} \\
\theta_6(A_iA_jB_jB_j) & = & \mu_{ij} + & (1+s-t)a_i + & (1-s+t)a_j + & (1-s-t+2st)d_{ij} \\
\theta_7(A_jA_jB_iB_i) & = & \mu_{ij} + & 2ta_i + & 2(1-t)a_j + & 2t(1-t)d_{ij} \\
\theta_8(A_jA_jB_iB_j) & = & \mu_{ij} + & (s+t)a_i + & (2-s-t)a_j + & (s+t-2st)d_{ij} \\
\theta_9(A_jA_jB_jB_j) & = & \mu_{ij} + & 2sa_i + & 2(1-s)a_j + & 2s(1-s)d_{ij}
\end{array}$$

where μ_{ij} is an unknown parameter representing a genetic background-dependent mean of the cross ij, a_i is the additive effect of the allele Q_i of the QTL and d_{ij} is the dominance effect between alleles Q_i and Q_j. s and t are functions of the recombination between markers and QTL defined by: $s=r_1r_2/(1-p)$, $t=r_1(1-r_2)/p$ and $p=r_1+r_2-2r_1r_2$, where r_1, r_2 and p are recombination rates between loci $A-Q$, $Q-B$ and $A-B$, respectively. Notice, that we assume absence of interference in meiotic recombination. If we suppose p known from the linkage map, we have only one parameter of position (*e.g.*, r_1) which we denote by x. Then, we get:

$$s = \frac{x(p-x)}{(1-p)(1-2x)} \quad \text{and} \quad t = \frac{x(1-p-x)}{p(1-2x)} \ .$$

For F_3 progenies we have the general model:

$$Y_{ijk.} = \sum_{l=1}^{9} \theta_l g_l + e_{ijk.} \ , \tag{1}$$

where $Y_{ijk.}$ is the phenotypic mean of F_3 individuals deriving from the k^{th} F_2 individual derived form the cross $L_i \times L_j$, $e_{ijk.}$ is the error term with variance σ^2, including environmental and other QTLs effects, g_l are dummy variables indexing the marker classes ($g_l = 1$ if the individual is from the class l and 0 otherwise).

Model (1) is linear and least squares could be used to estimate the parameters. We have a total of sixteen parameters where only eleven are estimable: 6 μ_{ij}, 3 a_i (*i.e.*, a_1, a_2 and a_3) and 2 d_{ij} (*i.e.*, d_{12} and d_{13}). So we have used the constraints:

$$\sum_{i=1}^{4} a_i = 0 \quad \text{and} \quad \sum_{j=1 \neq i}^{4} d_{ij} = 0 \quad \text{for each } i = 1..4 \ .$$

Model (1) can be written as: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, where $\mathbf{Y}$ is the $(N,1)$ vector of observations, $\mathbf{X}$ is the (N,r) matrix of the model, $\boldsymbol{\beta}$ is the $(r,1)$ vector of parameters and $\mathbf{e}$ is the $(N,1)$ vector of residuals supposed to have a normal distribution with mean $\mathbf{0}$ and variance $\sigma^2\mathbf{I}$. $\mathbf{X}$ and $\boldsymbol{\beta}$ can be decomposed as: $\mathbf{X} = [\mathbf{X}_0 \mid \mathbf{X}_1 \mid \mathbf{X}_2]$ and $\boldsymbol{\beta}^t = [\boldsymbol{\beta}_0^t \mid \boldsymbol{\beta}_1^t \mid \boldsymbol{\beta}_2^t]$ where $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are vectors of μ_{ij}, a_i and d_{ij}, respectively, and $\mathbf{X}_0$, $\mathbf{X}_1$ and $\mathbf{X}_2$ are the corresponding submatrices. Elements of $\mathbf{X}_0$ are 0 or 1 according to the cross to which the individual belongs and those of $\mathbf{X}_1$ and $\mathbf{X}_2$ are coefficients of the a_i and d_{ij} which are calculated at position x according to the marker interval considered. In these conditions the best linear unbiased estimators are ordinary least square ones given by: $\hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}$ and $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \mathbf{V} = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1})$ with $\hat{\mathbf{V}} = \hat{\sigma}^2(\mathbf{X}^t\mathbf{X})^{-1}$ and $\hat{\sigma}^2 = \mathbf{Y}^t(\mathbf{I} - \mathbf{X}\mathbf{X}^-)\mathbf{Y}/(N-r)$, where $\mathbf{X}^- = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$ and r is the number of estimated parameters.

At any position x (say every centiMorgan), the presence of a QTL can be tested through several hypotheses. Two of these are:

$$H_{01}: \forall\, i,j \quad a_i = d_{ij} = 0 \qquad T_1(x) = \frac{\mathbf{Y}^t(\mathbf{X}\mathbf{X}^- - \mathbf{X}_0\mathbf{X}_0^-)\mathbf{Y}}{\mathbf{Y}^t(\mathbf{I} - \mathbf{X}\mathbf{X}^-)\mathbf{Y}} \, \frac{N-r}{q_1}$$

H_{02}: $\forall\ i,j \quad a_i = 0$ assuming "$d_{ij} = 0$" when d_{ij} are not necessarily zero:

$$T_2(x) = \frac{\mathbf{Y}^t(\mathbf{X}_{01}\mathbf{X}_{01}^- - \mathbf{X}_0\mathbf{X}_0^-)\mathbf{Y}}{\mathbf{Y}^t(\mathbf{I} - \mathbf{X}_{01}\mathbf{X}_{01}^-)\mathbf{Y}} \, \frac{N-r}{q_2} \, ,$$

where $\mathbf{X}_{01} = [\mathbf{X}_0 \mid \mathbf{X}_1]$ and q_1 and q_2 are the numbers of tested parameters in T_1 and T_2, respectively. Notice, that $T_2(x)$ tests the hypothesis "$a_i = 0$" when "$a_i = d_{ij} = 0$" is true. This means that $T_2(x)$ tests only additive effects supposing that dominance is absent when it could be present (*cf.* Rebaï & Goffinet 1993). Under H_0, $T_1(x)$ and $T_2(x)$ are distributed as $F(5,N-r)$ and $F(3,N-r)$ respectively. As discussed in Rebaï & Goffinet (1993) T_2 could be more powerful than T_1 when dominance is small.

Genetic interactions could be tested using the same approach. Two of these are very interesting: additive by additive (AA) epistasis defined as the interaction effects between the a_i parameters of two QTLs and the genetic background interaction (BA) between the QTL effects a_i (possibly d_{ij}) and the means μ_{ij}. To simplify the task and to avoid working with cumbersome models we choose to test these effects with individual marker models. Instead of considering the QTL itself we consider the most closely linked marker(s) to the QTL. Inferences about interaction effects between pairs of independent markers or between a marker and the means could then be easily done. For instance, for the BA interaction in F_3, we have the model for a marker M:

$$\begin{aligned}
Y_{ijk.} &= \mu_{ij} + 2a_i + 2\gamma_{iji} + e_{ijk.} && \textit{if} \quad G(k) = M_iM_i \\
Y_{ijk.} &= \mu_{ij} + a_i + a_j + d_{ij} + \gamma_{iji} + \gamma_{ijj} + e_{ijk.} && \textit{if} \quad G(k) = M_iM_j \\
Y_{ijk.} &= \mu_{ij} + 2a_i + 2\gamma_{iji} + e_{ijk.} && \textit{if} \quad G(k) = M_jM_j
\end{aligned}$$

where $G(k)$ is the genotype of the F_2 individual k for marker M and γ_{iji} is the interaction $\mu_{ij}{*}a_i$. There are 12 parameters γ in all but only one is estimable. The model for epistasis implies two markers and is more complicated. For these interactions, suitable constraints were chosen to determine the models and tests were calculated as for a_i and d_{ij}. When testing at a global significance level α, a per-test level α/n_t should be used, where n_t is the number of tests performed (supposed independent). For example, if we test, at a 5% level, epistasis between 5 independent QTLs we perform 10 tests and we should take a 0.5% level for each test. A sequentially rejective multiple test procedure (Holm 1979) could also be used.

In QTL mapping experiments one would be interested to know if the same QTL is detected in two different environments by testing whether the apparent position of a QTL differs between them. Stuber *et al.* (1992) proposed to compare the maxima of test statistics obtained in each environment to that obtained when analysing the two environments together. However, a simple way to test the QTL by environment interaction is to use an analysis of variance model involving three factors: the population *ij*, the marker genotype, the environment and their interactions.

Implementation of the method

QTL detection

The test statistic $T(x)$ (T_1 or T_2) is performed every centiMorgan for each chromosome and a QTL is declared when this statistic exceeds a predetermined threshold. The likely position of the QTL is then that corresponding to the maximum of $T(x)$ values. However, as tests $T(x)$ on a chromosome are statistically correlated, the threshold at a level α for the global test T (defined as $Sup\,(T(x),\ 0 < x < L)$, L is the chromosome length), is not easy to obtain. In Rebaï *et al.* (1994b) we have proposed analytical and simulation-based approximations to get appropriate thresholds in a large number of situations. These were used to calculate thresholds for T_1 and T_2 at a per-chromosome level $\alpha = 1\%$ (Table 1).

Once a QTL is located, parameters of this QTL could be estimated on its likely position. A way to express the global effect of the QTL is to calculate:

$$\hat{\sigma}_a^2 = \sum_{i=1}^{4} \hat{a}_i^2 , \quad \hat{\sigma}_d^2 = \sum_{i=1}^{4} \sum_{j=1 \neq i}^{4} \hat{d}_{ij}^2 \quad \text{and} \quad \hat{\sigma}_q^2 = \hat{\sigma}_a^2 + \hat{\sigma}_d^2 ,$$

which represent additive, dominance and total variance due to the QTL, respectively. The effect, evaluated in % of the phenotypic variance, could then be calculated by $R^2 = 100\,\hat{\sigma}_q^2 / (\hat{\sigma}_q^2 + \hat{\sigma}^2)$. This quantity would be a biased estimator of the true coefficient of explication of the QTL, especially because it is calculated at the estimated position of the QTL which is not a consistent estimator of the actual position. Support intervals (SI) for the likely position of the QTL could be calculated as proposed by Lander & Botstein (1989). One takes the positions at which test values are one unit below the maximum to be the limits of the support interval. This procedure gives confidence intervals for the QTL position at 60-95% levels depending on the QTL effect and the marker density

(Mangin *et al.* 1994).

Programming and simulations

The method described above was programmed under the Interactive Matrix Language SAS/IML (1985). The marker genotypes within each population were denoted by 1 and 2 for the homozygotes, 3 for the heterozygotes and 0 for missing data or when the marker is not polymorphic in the population considered. This last case happens for markers having less than 4 alleles in the original parents. For each individual, coefficients of a_i and d_{ij} are calculated given the genotypes of the flanking markers. When one of these markers is 0 we use the closest informative one (not 0). If the 0 marker is the first or the last on the chromosome, expectation of such individuals are obtained by using only the three classes of the closest non 0 marker. [Genotypes of the bordering markers could also be inferred from those of the other markers.] For each position, the elements of matrix **X** are calculated line by line. Computations to obtain the test values and the estimates are then straightforward.

Programs were written for both F_3 and *TC* progenies. They were first tested on simulated data with 100 individuals in each F_2 population (600 in all). A chromosome of 100 cM, having 6 equidistant markers and a QTL at 50 cM with $R^2 = 0.10$ and equal additive and dominance variances, was simulated. Parameters were: $\mu_{ij} = 0$, $a_1 = -a_3 = 0.15$, $a_2 = -a_4 = 0.30$, $d_{12} = d_{34} = 0.66$, $d_{13} = d_{14} = d_{23} = d_{24} = -0.33$ ($\sigma_a^2 = \sigma_d^2 = 0.055$) and $\sigma^2 = 1$. Markers 1, 3 and 6 have 4 alleles, markers 4 and 5 have 3 alleles and marker 2 only 2 alleles.

Results (Table 1) show that the QTL is consistently detected with our tests in both progenies. Thresholds and empirical estimates of the power (*i.e.*, the ratio of replicates that are significant) are also presented. The tests have good powers and the QTL position is well estimated. R^2 estimates have large standard errors and have a small bias in *TC* progenies.

Table 1. Tests T_1 and T_2 for F_3 and *TC* applied to simulated data[a]

Test	T[b]	Power	Position[c]	$\hat{R}^2$ (%)	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	$\hat{a}_4$	$\hat{d}_{12}$	$\hat{d}_{13}$	
T_1 F_3	3.90	0.80	47 (39-55)	12.2	0.14	0.27	-0.16	-0.25	0.66	-0.21	
T_2 F_3	5.00	0.85	46 (30-62)	*5.1*	*0.06*	*0.07*	*0.08*	*0.05*	*0.21*	*0.12*	[d]
T_1 *TC*	4.10	0.95	49 (43-55)	14.3	0.21	0.29	-0.21	-0.23	0.46	-0.24	
T_2 *TC*	5.20	0.97	51 (29-73)	*1.2*	*0.03*	*0.07*	*0.06*	*0.03*	*0.08*	*0.06*	[d]

[a] 100 replicates with a population size of 600. [b] thresholds at 1% level. [c] empirical confidence intervals at 95%. [d] empirical standard deviations of estimates in italics

Table 2. QTLs for days to silking (DS) and ear height (HE) in F_3 and *TC* progenies

Character		Chr	Position[a]	$\hat{R}^2$ (%)	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	$\hat{a}_4$	$\hat{d}_{12}$	$\hat{d}_{13}$
DS	F_3	2	90 (70-98)	14.2	0	0	0.50	-0.55	1.24	1.36
	TC	2	4 (0-14)	4.2	-0.23	0.76	-0.31	-0.22	-0.16	0
DS	F_3	4	0 (0-20)	4	0.87	0	0	-0.5	0	-0.25
	TC	4	7 (0-24)	4.2	0.41	-0.18	0.35	-0.6	0.2	-0.1
HE	LO[b]	3	0 (0-30)	4.4	0.11	0	-0.19	0.14	-0.09	0.08
	LT	3	22 (4-50)	4.0	0	0	-0.31	0.22	-0.08	0
HE	LO	4	29 (19-33)	7.6	0.25	0	0	-0.18	0.18	-0.10
	LT	4	27 (16-34)	8.1	0.23	0	0	-0.19	0.19	-0.11

[a] with support intervals. [b] LO and LT are two different locations

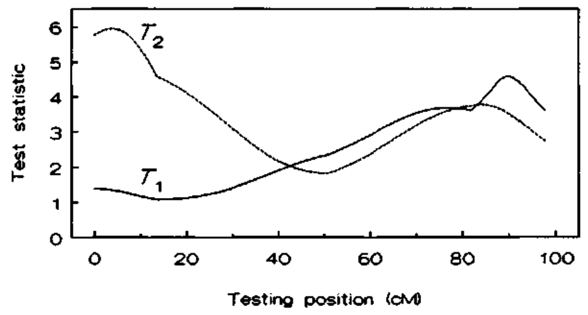
Application to experimental data

The method was applied to a diallel cross among four inbred lines of maize (*Zea mays* L). The six F_2 populations (800 individuals in all) were genotyped with respect to RFLP and isozymes markers and the linkage map established using JoinMap (Stam 1993). *TC* and F_3 were evaluated for several characters in two and one location(s), respectively. For *TC*, 1500 individuals were measured in each location.

Results for days to silking (DS, measured in days from sowing to ear silk emergence in both progenies) and ear height (HE, height to ear node in dm) for chromosomes 2, 3 and 4 are presented in Table 2. A postulated QTL was detected by significance of T_1 and/or T_2. The QTLs found have individual effects ranging from 4 to 14% of the phenotypic variance and are of different natures.

For DS, QTLs detected on F_3 were also found in *TC* with different effects and positions. For F_3, QTL mapping was also carried out within populations. Results (not shown) are in good agreement with the allelic effects globally estimated, *i.e.*, that when $\hat{a}_i - \hat{a}_j$ is quite large, the QTL is detected in population *ij*. Except for chromosome 2 (Figure 1), where it is likely that two different QTLs are identified in F_3 and *TC*, the

Figure 1. Tests T_1 in F_3 and T_2 in *TC* for days to silking on chromosome 2



QTLs are globally consistent among progenies. In F_3 the QTL was detected only by T_1 and has a large dominance effect, whereas in TC the QTL has a large additivity and was detected by T_2. For HE, most QTLs were detected in both environments with a good consistency of positions and effects.

Epistasis (AA) and interaction with the common genetic background (BA) were investigated for detected QTLs and were found non significant at the global 5% level.

Acknowledgments

This work was partly supported by Rustica semences, Caussade semences and Maïs Adour.

References

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

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6: 65-70.

Knapp, S.J., W.C. Bridges & D. Birkes, 1990. Mapping quantitative trait loci using molecular marker linkage maps. Theor Appl Genet 79: 583-592.

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

Mangin, B., B. Goffinet & A. Rebaï, 1994. Constructing confidence intervals for QTL location. Genetics: submitted.

Rebaï, A. & B. Goffinet, 1993. Power of tests for QTL detection using replicated progenies derived from a diallel cross. Theor Appl Genet 86: 1014-1022.

Rebaï, A., B. Goffinet & B. Mangin, 1994a. Comparing power of different methods for QTL detection. Biometrics: in press.

Rebaï, A., B. Goffinet & B. Mangin, 1994b. Approximate thresholds of interval mapping tests for QTL detection. Genetics 138: 235-240

SAS/IML User's Guide, Version 5, 1985. SAS Institute Inc. Cary, North Carolina, USA.

Stam, P. 1993. Construction of integrated genetic linkage maps by means of a new computer package: JoinMap. The Plant Journal 5: 739-744.

Stuber, C.W., S.E. Lincoln, D.W. Wolff, T. Helentjaris & E.S. Lander, 1992. Identification of genetic factors contributing to heterosis in a hybrid from two elite maize inbred lines using molecular markers. Genetics 132: 823-839.

Estimates of relationships between quantitative traits and molecular markers by means of Genetic Classifier Systems

Federico Mattia Stefanini & Alessandro Camussi, Genetic Unit, Agricultural Faculty, University of Florence, via S. Bonaventura 13, 50145 FIRENZE, ITALY

Key words
genetic classifier systems, genetic algorithms, molecular markers

List of abbreviations
GCS = genetic classifier system, GA = genetic algorithm

Abstract

Linear models are currently used to identify individual loci responsible for the expression of quantitative traits and to predict phenotypic values from molecular profiles.

Recent development of molecular technology and the consequent availability of a large number of markers can impose severe limitations on the use of linear models.

The increasing complexity of genetic information requires, for this reason, new analytical tools. Among others, Genetic Classifier Systems can be a promising approach, being adaptive algorithms which deal with complex problems in an easy but powerful way.

A simple Genetic Classifier System is presented and applied to a data set of reduced dimension. The results are compared with those obtained by application of a logistic model.

Introduction

The potential use of molecular markers in breeding programs are mainly 1) fingerprinting and determination of relationships between individuals, and 2) identification and selection of favourable genes, on the basis of existing linkage with markers. Different classes of markers are, at present, available in most laboratories. The possibility to increment the number of useful polymorphic markers is now a reality, because of laboratory automation and the availability of several hundreds of public RFLP probes, at least in extensively studied species.

Several statistical methods, mainly based on linear models, have been developed to detect the association between genetic markers and favourable genes, to provide unbiased estimates and to minimize the possibility of false assignments (see Jansen; Utz and Melchinger; Hackett, in these proceedings). The use of linear models can reveal a number of limitations, particularly if a weak association is present among many different markers, or in the presence of missing observations.

We suggest to integrate the approach based on linear models with the use of Genetic Classifier Systems. These systems comprise a class of adaptive machine learning systems built on three functional components: 1) a "rule and message" system, 2) an "apportionment of credit" system, 3) a genetic algorithm (GA) (Holland 1975, 1987, 1992; for a mathematical description: Holland 1986). Within a Genetic Classifier System (GCS) no explicit criterion is formulated about how to achieve the goal, but the algorithm itself learns by an iterated procedure from the set of the available experimental data, and on the basis of a rewarding system. Thus, GCSs could be a convenient tool in the study of complex systems.

General features of GCSs

The "rule and message" system is a way to implement relations of the type <IF "condition" THEN "action">. These condition and action are parts of each rule and they are coded using strings of fixed length on a finite alphabet, like {0,1,#}. It is proved that the rules system is computationally complete and represents compactly the knowledge basis.

The string based representation of applied problems allows the search of better candidates in the space of admissible rules using a genetic algorithm (*e.g.*, Stefanini & Camussi 1993).

A GA is a system that simulates a natural population where the selection occurs. Each individual (a rule) has a probability of staying into the next generation directly related to its attitude in specifying an actual relation between the action-part and the condition-part it expresses. An "apportionment of credit" system numerically quantifies the usefulness of rules (their fitness) while they are working towards the goal. New rules are introduced into the population mainly by means of crossing-over and mutation operators that are analogues to the biological phenomena.

At the start of a computer run, each rule has the same fitness value so that a flat fitness landscape is defined. At the end of the run, the landscape can show many peaks representing a set of co-adapted rules that constitute the information core extracted from the experimental data set.

A rule is the smallest unit of information the GCS works on. It is typically represented by two strings separated by a slash, *e.g.*, 10010 / 0110. A rule is a compact notation for an input/output relation. The set of input values and the set of output values are labelled by binary strings and the rule 10010 / 0110 can be translated as " if the input value is 10010 then set the output value to 0110".

Let A="10010 / 0110" be a rule taken as an example. If we ignore the single character in a generic position *j* of the condition-part, then two values are indicated at the same time; that is, the two values differ only at the position *j*. We indicate this by putting the character "#", usually called "don't care" symbol, into position *j*. The same procedure can be performed in the action part to indicate two different output values at the same time. Therefore the single rule can establish a relation expressed by the link of a subset of input values with a subset of output values.

A general input/output relation is obtained by a collection of rules of this sort, where several positions are set up with "don't care" symbols.

The relative *specificity* of a rule is the ratio between the length of the specified (by 0 or 1) part and the total length of a rule.

The power of GCSs to deal with complexity has been shown by many successful applications since the '80s. Some examples are: systems for medical diagnosis, morphogenesis simulations, prediction of company profitability, description of consumer preferences, and a system of gas pipeline control (for more details see Goldberg 1989, chapter 6).

GCS in the study of relationships between markers and measured traits

We suggest to consider a GCS like a black box that receives molecular information as input from experiments and that gives, as output, the forecast trait of interest. A rule specifies a relation between a subset of all theoretical genotypes (the action-part) and a subset of trait values (the condition-part), also if multivariate. More formally, input and output values belong to the Cartesian product of as many sets of attributes as the variables are.

If a rule specifies an actual relation, its fitness parameter will be improved along with the simulation running.

Let us consider some simple cases of relations among marker genotypes and trait values. Let M1, M2, M3 be three markers (coded as: 0=presence, 1=absence) and V1 a binary trait of interest (*e.g.*, 0=susceptible, 1=tolerant). If the marker status M1=0 is always related to trait values V1=0 and if marker status M1=1 is related to V1=1 and M2, M3 are independent, then only two rules R are requested to sum up the overall information, namely R1="0## / 0" and R2="1## / 1".

Now, suppose that M1, M2, M3 are the same markers and V1 is a four attributes trait value (coded as 00,01,10,11). If marker status M2=0 is always related to trait value V1=01 or V1=00 and if M2=1 is related to V1=10 or V1=11, and M1, M3 are independent, then the requested rules are R1="#0# / 0#" and R2="#1# / 1#".

The distribution of trait values, when a fixed set of marker attributes is considered, determines which type of rules will result as co-adapted.

The computer simulation realizes a sampling procedure that will improve the fitness of rules allowing right predictions within the class of more frequent molecular inputs. These rules become members of the final cluster because they are more frequently rewarded.

The set of co-adapted rules can be used to predict individual trait values, to chose a reduced number of meaningful markers or to exclude the set of redundant ones.

A critical point in real applications is due to thousands of markers and hundreds of trait classes whose relation is unknown. It is unrealistic to extract this type of information without an optimal computerised algorithm. An example of a general purpose algorithm is explained in Holland (1986).

The algorithm

In this paper a simple GCS, derived from Goldberg's genetic based machine learning (1989), is considered. It was implemented using Borland C++ compiler for PC-DOS.

A computer run is constituted mainly by the following cycle: 1) to take a message-input from the "environment" (the experimental data set); 2) to compile a list of rules whose conditions are satisfied, *i.e.*, all messages that match their condition-part; 3) to choose one rule within the list of matched rules, called "winner"; 4) to verify if the action-part of the winner is in accordance with the observed trait value of the current input; and to increase its fitness if it has made the right prediction.

Two arrays are present, one to store a population of rules, the other to store fitness, specificity and ancillary information.

The initialisation routine mainly creates a list of random rules, and it sets up the starting values of fitness and specificity. These random rules are typically of the dimension 50 up to 1000, a small dimension in comparison with all possible theoretical rules.

The main cycle described above is repeated thousands of times to obtain an optimised fitness landscape where good rules survive and bad rules die off.

The competition among matched rules (given an input) produces a winner that is characterised by the best "bid value". The bid value is established as a linear function of the rule's fitness and specificity. At each main cycle the fitness of each rule is also decreased by a fixed amount, called "life tax", to minimise the presence of bad, unrewarded rules over generations.

The bid made by a rule i at cycle t is (Goldberg 1989, modified):

$$B_i(t) = (C_{bid} + C_{spe})\cdot S_i(t) \ ,$$

where $S_i(t)$ indicates the fitness, and C_{bid} and C_{spe} are constants. The difference equation of fitness changes is:

$$S_i(t+1) = S_i(t) - P_i(t) - T_i(t) + R_i(t) \ ,$$

where $T_i(t)$ is the "life tax" amount that is proportional to the fitness, $R_i(t)$ is the reward if the right marker-trait relation is established by the rule, and, accordingly, $P_i(t)$ is equal to $B_i(t)$ if the rule i is the winner at cycle t, otherwise it is null.

At this step no new rule is introduced into the list, but only a scoring system is available. The GA is invoked after a fixed number of cycle repetitions. It draws out two rules using fitness as probability and eventually changes them by means of crossing-over and mutation operators. A crossing-over operator copies strings with substring exchanges, and a mutation operator randomly changes single bits, *e.g.*, 0 becomes 1 or #. The place where mutation occurs and the type of bit change are randomly assigned.

Particular care is reserved to the choice of rules substituted in the next cycle. A scheme similar to the proposal of De Jong (1975) is applied, based on the assumption that the optimal choice is represented by a switch between the new rule and an old rule highly similar to it and characterised by a small fitness value.

We propose to include a "profile of univariate expectation", $U_i(t)$ to obtain an optimised bid, B_i, to improve GCS performances in molecular studies. It is a weighted linear summation that quantifies the amount of highly informative markers that a rule specifies, namely whose value is assigned equal to 0 or 1. If a marker is really the gene of a trait, or if it is highly linked to it, then it has the maximum amount of information, because its knowledge is enough to establish the trait value of interest. A monomorphic marker has the minimum amount of information, providing no prediction about the trait. Molecular markers usually have an intermediate amount of information.

In the case of a binary trait, $U_i(t)$, depends on the statistic u_j, for marker j, defined as:

$$u_j = \frac{\left| \sum \mathrm{I}_{\{x=1 \cap phenotype=0\}}(x) - \sum \mathrm{I}_{\{x=1 \cap phenotype=1\}}(x) \right|}{\max\left[\left| \sum \mathrm{I}_{\{x=1 \cap phenotype=0\}}(x) - \sum \mathrm{I}_{\{x=1 \cap phenotype=1\}}(x) \right| \right]} ,$$

with $\mathrm{I}_{\{.\}}(x)$ the characteristic function indicating when the argument of the summation is not null. Thus, if marker j is considered, then the difference between observations with marker allele one within the trait value zero and the observations with marker allele one within the trait value one quantifies the information that marker j has about the trait value, with the denominator introduced as normalising term.

During a computer run, a generic rule i at cycle t has a "profile of univariate expectation" $U_i(t)$ equal to:

$$U_i(t) = \frac{\sum u_j * \mathrm{I}_{\{1,0\}}(x)}{m} ,$$

with j the marker index, m the total number of markers, $\mathrm{I}_{\{.\}}(x)$ the characteristic function taking into account only markers to which the "don't care" symbol is not assigned. Thus, $U_i(t)$ will improve the fitness of rules with markers of high univariate expectation.

A case study: a data set of ten individuals with a two valued trait and 30 binary markers
The performances of the proposed GCS were assessed in very stringent conditions (a data set with more markers than individuals) and compared with the results obtained by logistic regression. The logistic regression analysis is indicated to investigate the relationship between the probability of a binary response variable and the explanatory variables (Andersen 1991). Different procedures are applied to search for a model: backward, forward and stepwise. Statistical calculations were made using SAS package, procedure "logistic" (SAS/STAT User's Guide, vol. 2, SAS Institute, Cary, USA).

The artificial data set was constituted by only 10 individuals characterised by a binary trait and 30 binary markers (scored as 0=presence, 1=absence). The marker M16 is completely correlated with trait values, while the remainders are randomly assigned.

The results of logistic regressions are reported in Table 1. It is evident that the presence of marker M16, fully correlated with the trait of interest, is only revealed by the forward procedure if convergence parameters are not stringent.

The most important features of the GCS output are reported in Table 2. The best eight

Table 1. Results from fitting a logistic model to the data set. The forward, the backward and the stepwise procedures were used to identify relevant markers with different values of MAXITER (maximum number of iterations) and CONVERGE (minimum amount of variation recognised at convergence)

PROCEDURE	MAXITER = 25
	CONVERGE = $1{*}10^{-4}$
		MAXITER = 1000
		CONVERGE = $1{*}10^{-2}$

Forward	Convergence was not attained	INTERCEPT = 101.2, M16 = −202.4
Backward	Convergence was not attained	INTERCEPT = 1.3863, M2 = −2.7725
Stepwise	Convergence was not attained	INTERCEPT = 0

rules, as regards their fitness values, are listed, along with related statistics. Because eight rules are needed to classify ten individuals, an information "gain" of 20% is obtained, even if the dimension of the data set is small. Figure 1 identifies the most relevant markers: M2 and M16 received 100% of assignment. It is interesting to note that the same markers were revealed by the logistic model even if using different procedures of model selection.

Discussion

The GCS gives valuable results in a data set of small dimension, where usual linear models failed to work correctly, at least with current settings of selection criteria. This new class of "machine learning" systems is also expected to work better in large data sets in which no simple correlation pattern is present among single molecular marker scores. Correlation among markers is accounted for by the algorithm, if an appropriate "bid" function is used. Relevant results are expected with GCSs if some requirements are fulfilled: 1) a relationship between molecular markers and quantitative traits exists; 2) these "regularities" are accessible to GCS exploration by means of an effective coding

Table 2. Main results from a G.C.S. run: the best 8 rules are reported. R=condition-part, P=action-part (*i.e.*, the coded trait values), F=fitness values, S=specificity, M=number of matchings. All ten marker genotypes are correctly matched with their trait values

R	P	F	S	M
#1001####0#####1#1#0#####0##0#	1	37.4144	0.3334	2
#0##0#0##0#####0#0##0#1#0##0##	0	33.4901	0.3334	2
#1####00#######0#0#######0##0#	0	33.3518	0.2334	1
#1####00######11###1####1#10##	1	32.5916	0.3000	1
#10#1####0##0##1#1#####1#1##1#	1	31.5854	0.3334	1
#0####00#######0#0##0#1#0##0##	0	31.0443	0.3000	1
#0##0#0##0#####1#1#1#####1##1#	1	30.5718	0.3000	1
#0##0####0#####0####1###110#0#	0	30.4824	0.3000	1

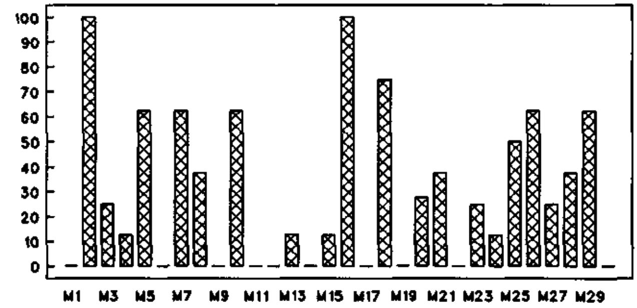


Figure 1. Percentage of marker assignment by 8 final rules. This statistic is related to the marker ability to predict trait values

system; 3) an optimised set of algorithmic parameters is set up; this point can be fulfilled using descriptive statistics about performances of computer runs.

The procedure outlined can be extended to true continuous traits. Because the finite resolution of measurement instruments and the finite sample dimension, discrete-qualitative recordings can be made (Rosen 1978). The extension of the procedure in this direction is in progress.

The proposed procedure is to be seen as an explorative tool to be used along with usual linear models with the aim to extract from a very large mass of information its relevant core; this core can be fully explored by statistical techniques allowing a deeper analysis of the relationships between molecular markers and traits of interest.

References

Andersen, E.B., 1991. The statistical analysis of categorical data, 2nd ed., Springer-Verlag, Berlin.

De Jong, K.A., 1975. An analysis of the behaviour of a class of genetic adaptive systems, (doctoral dissertation, University of Michigan), Dissertation Abstracts International, Michigan, 5140B, 36(10).

Goldberg, D.E., 1989. Genetic algorithms in search, optimization and machine learning. Addison Wesley, New York.

Holland, J.H., 1975. Adaptation in natural and artificial systems. Ann Arbor, The University of Michigan Press, Michigan.

Holland, J.H., 1986. A mathematical framework for studying learning in classifier systems. Physica 22D: 307-317.

Holland, J.H., 1986. Escaping brittleness: the possibility of general purpose learning algorithms applied to parallel rule based systems. In: R.S. Michalski, Carbonell, J.G. & T.M. Mitchell (Eds.), Machine learning II, pp. 595-623. Morgan Kaufmann, Los Altos CA.

Holland, J.H., 1987. Genetic algorithms and classifier systems: foundations and future directions, In: J.J. Grefenstette (Ed.), Genetic algorithms and their applications: Proceedings of the second international conference on genetic algorithms, pp. 82-89. Lawrence Erlbaum Associates Publishers, London.

Holland, J.H., 1992. Genetic Algorithms, Sci Amer 267(1): 44-51.

Rosen, R., 1978. Fundamentals of measurements and representation of natural systems. North-Holland. New York.

Stefanini, F.M. & A. Camussi, 1993. APLOGEN: an object oriented Genetic Algorithm performing Monte Carlo optimization. CABIOS 9: 695-700.

Strategies of pooling for parentage analyses applying DNA markers

Jussi Tammisola[1], Satu Åkerman[1], Mikko Regina[2], Seppo Lapinjoki[1] & Veli Kauppinen[1],
[1] VTT, Biotechnology and Food Research, P.O. Box 1505, FIN-02044 VTT, Espoo,
[2] Dept. of Pharmacy, Univ. of Kuopio, P.O. Box 1627, Kuopio, Finland

Key words
contamination, genetic markers, pooling, populations, progeny mixtures

Abstract

Pooling can be efficiently utilized in determining parental and progeny relationships in natural populations and practical breeding. As a consequence, the number of necessary PCR reactions and/or DNA extractions is reduced. Parents can be determined by applying several parental candidate pools or a single progeny pool, and progeny individuals can be identified by studying progeny candidate pools. Pooling can also be applied to determining the parents of a progeny mixture, and to screening for progeny contamination. The efficiency of identification in relation to pool size, DNA marker types, and non-distinguishable alleles is discussed.

Introduction

Mating structure, distribution of propagules, and other aspects of population structure can be studied in natural populations by applying DNA markers. Parentage analyses are also needed for quality control in breeding, when the supposed genetic origins of the most valuable materials are checked. Mistakes are known to occur due to *e.g.* mixing or mislabelling of samples, base stem escape in grafting, pollen contamination and coding errors. However, large scale parentage analyses may prove expensive, because the extraction of plant DNA is often laborious, and PCR enzyme costs are high. In this work it is shown that these obstacles can be largely overcome by using pooled materials in the analyses. Related ideas have proved successful in studying linkage (Michelmore *et al.* 1991, Williams *et al.* 1993b, Taylor *et al.* 1994) and genetic relatedness of populations (Yu & Pauls 1993). The first practical application of these principles in parentage analysis was a study in European white birch (*Betula pendula* Roth) using RAPD markers (Åkerman *et al.* 1994).

Determining parents by applying parental candidate pools

The classical analysis of parentage by eliminating impossible parental combinations (*e.g.*, Ellstrand 1984) is generalized for parental candidate pools. In order to utilize the available information more completely, pool combinations are eliminated instead of pools. The two true parents of an individual are searched from a large set of parent candidates by consecutive cycles, in which

(a) the remaining parental candidates are combined into several pools,
(b) marker phenotypes of the individual and the pools are determined,
(c) the pool combinations that cannot have produced the individual are always eliminated (Figure 1).

From now on, a diploid and cross-breeding species with codominant and polyallelic markers is assumed, unless otherwise stated. Denoting the marker genotype of the individual in concern by a_i/a_j, the probability of elimination (P_e) of a pool combination ($r \times s$) is

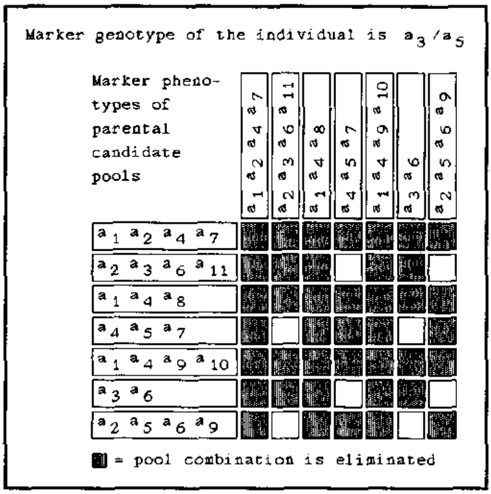


Figure 1. Elimination of the parental candidate pool combinations which cannot have produced the individual. A diploid species and codominant markers are assumed

Marker genotype of the individual is a_3/a_5

Marker phenotypes of parental candidate pools	$a_1 a_2 a_4 a_7$	$a_2 a_3 a_6 a_{11}$	$a_1 a_4 a_8$	$a_4 a_5 a_7$	$a_1 a_4 a_9 a_{10}$	$a_3 a_6$	$a_2 a_5 a_6 a_9$
$a_1 a_2 a_4 a_7$	■	■	■	■	■	■	■
$a_2 a_3 a_6 a_{11}$	■	■	■	□	■	■	□
$a_1 a_4 a_8$	■	■	■	■	■	■	■
$a_4 a_5 a_7$	■	□	■	■	■	□	■
$a_1 a_4 a_9 a_{10}$	■	■	■	■	■	■	■
$a_3 a_6$	■	■	■	□	■	■	□
$a_2 a_5 a_6 a_9$	■	□	■	■	■	□	■

■ = pool combination is eliminated

$$P_e\{r \times s\} = \frac{1}{\binom{N}{2N_p}\binom{2N_p}{N_p}} \cdot \left\{ \binom{N-n_i}{N_p}\binom{N-N_p-n_i}{N_p} \right. + \binom{N-n_j}{N_p}\binom{N-N_p-n_j}{N_p} + \binom{N-n_{i+j}}{N_p}\binom{N-N_p-n_{i+j}}{N_p} \left. + 2 \cdot \binom{N-n_{i+j}}{N_p} \cdot \left[\binom{N-N_p}{N_p} - \binom{N-N_p-n_i}{N_p} - \binom{N-N_p-n_j}{N_p} \right] \right\}, \tag{1}$$

in which N_p = pool size, n_ℓ = number of parental candidates containing allele a_ℓ ($\ell = i,j$) and n_{i+j} denotes the total number of parental candidates containing either allele a_i or a_j. Because both of the true parents may be situated in a common pool, the probability of elimination is also calculated for the combination of a parental candidate pool r with itself ($r \times r$), which equals

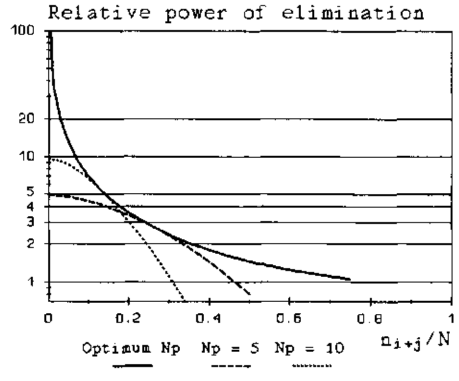
$$P_e\{r \times r\} = \frac{\binom{N-n_i}{N_p} + \binom{N-n_j}{N_p} - \binom{N-n_{i+j}}{N_p}}{\binom{N}{N_p}} . \tag{2}$$

Therefore, for a single marker locus, the probability of elimination of a parental candidate combination ($t \times u$) is

$$P_e\{t \times u\} = \frac{1}{(N-1)} \cdot \left[(N-N_p) \cdot P_e\{r \times s\} + (N_p-1) \cdot P_e\{r \times r\} \right] . \tag{3}$$

The number of PCR reactions required at this marker locus is $1 + N/N_p$. The respective figures without pooling are obtained by substituting 1 for N_p in the formulas. Hence, the relative power of elimination (RP_e) of parental candidate combinations ($t \times u$) per PCR reaction with pooling of parental candidates compared to without pooling is

Figure 2. Relative power of elimination (RP_e) of parental candidate combinations per PCR reaction with applying parental candidate pools compared to without applying pools. Without pooling $RP_e = 1$. The marker genotype of the progeny individual considered is a_i / a_j, and n_{i+j} / N = proportion of parent candidates containing either progeny allele. N = No. of parent candidates (N = 200), and N_p = pool size. Conservative examples, in which $n_{i+j} = n_i + n_j$

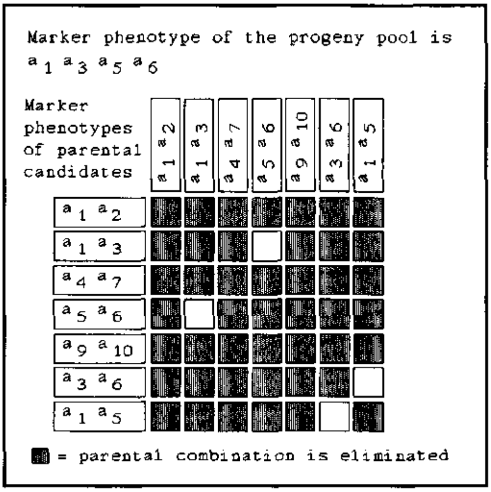


$$RP_e\{t\times u\} = \frac{N_p(N+1)}{(N+N_p)(N-1)} \cdot \frac{(N-N_p)\cdot P_e\{r\times s\} + (N_p-1)\cdot P_e\{r\times r\}}{\left[1 - \frac{2n_i n_j - n_{ij}(n_{ij}+1)}{N(N-1)}\right]} . \tag{4}$$

Here n_{ij} = No. of parent candidates containing both progeny alleles (a_i and a_j).

In terms of PCR reactions, the true parents of an individual can often be found in a more efficient way by studying parental candidate pools rather than the candidates separately (Figure 2). Pooling proves especially favourable when the marker alleles of the individual in concern are rare within the population of its parental candidates.

Figure 3. Determining parents by applying a progeny pool. A parental candidate combination is eliminated, if the progeny pool is missing any of its marker alleles. A representative pool size and no contamination are assumed



Marker phenotype of the progeny pool is $a_1\ a_3\ a_5\ a_6$

Marker phenotypes of parental candidates	$a_1 a_2$	$a_1 a_3$	$a_4 a_7$	$a_5 a_6$	$a_9 a_{10}$	$a_3 a_6$	$a_1 a_5$
$a_1\ a_2$	▨	▨	▨	▨	▨	▨	▨
$a_1\ a_3$	▨	▨	▨	□	▨	▨	▨
$a_4\ a_7$	▨	▨	▨	▨	▨	▨	▨
$a_5\ a_6$	▨	□	▨	▨	▨	▨	▨
$a_9\ a_{10}$	▨	▨	▨	▨	▨	▨	▨
$a_3\ a_6$	▨	▨	▨	▨	▨	▨	□
$a_1\ a_5$	▨	▨	▨	▨	▨	□	▨

▨ = parental combination is eliminated

Determining parents by applying a progeny pool

With only a few parent candidates, pooling may not be worthwhile, but the progeny individuals can be pooled together instead. Marker loci are studied one by one, and their allele constitution is recorded in each individual parent candidate and in the progeny pool. All pairwise parental candidate combinations which cannot have produced the progeny pool are always eliminated (Figure 3).

Irrespective of whether pooling is applied or not, the number of progeny individuals analysed (= pool size N_p) has to be sufficiently large for assuring that every parental marker allele is represented in the progeny sample. Assuming that four marker alleles occur in a true pair of parents, an upper limit for the risk probability (α_e) of incorrectly eliminating the parental combination due to sampling error in the progeny is

$$\alpha_e \leq 1 - \left[1 - 2\cdot\left(\frac{1}{2} + \frac{1}{2}\cdot p_c\right)^{N_p} + p_c^{\,N_p}\right]^2 , \qquad (5)$$

in which p_c denotes the proportion of contamination in the progeny. Assuming no contamination, the pool sizes of 7, 9 and 12 progeny individuals are adequate at the risk probabilities of 0.05, 0.01 and 0.001, respectively (Table 1).

The relative power of elimination of parental candidate combinations ($t \times u$) per PCR reaction with pooling of progeny compared to without pooling is

$$RP_e\{t \times u\} \leq \frac{N + N_p}{N + 1} \, . \qquad (6)$$

This figure is an upper limit, because the probability of elimination (P_e) of parental candidate combinations is slightly lower with than without pooling. Pooled-progeny analysis cannot discriminate *e.g.* between the parental combinations $a_i / a_j \times a_k / a_l$ and $a_i / a_k \times a_j / a_l$, whereas either combination may be eliminated if the progeny individuals

Table 1. Number of progeny (N_p) to be analysed in order to prevent an incorrect elimination of the true parent or parental combination due to sampling error. The parent or parental combination is eliminated whenever the progeny pool is missing any of its marker alleles. Four codominant alleles are assumed in the combination

Proportion of contamination p_c	Risk probability for an erroneous elimination of a single parent			Risk probability for an erroneous elimination of a parent combination		
	0.05	0.01	0.001	0.05	0.01	0.001
0	6	8	11	7	9	12
0.10	7	9	13	8	11	14
0.30	9	13	18	11	14	20
0.50	13	19	27	16	21	29

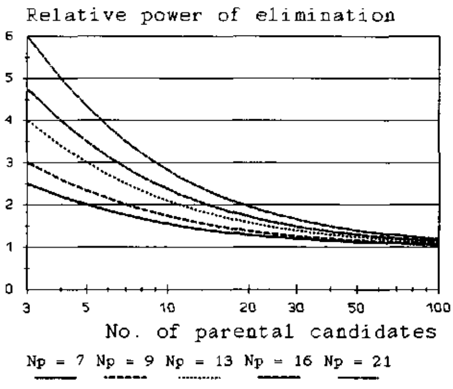


Figure 4. Relative power of elimination (RP_e) of parental candidate combinations per PCR reaction with applying a single progeny pool compared to without applying pooling. Without pooling $RP_e = 1$. N_p = No. of progeny individuals analysed (= pool size). Low marker allele frequencies are assumed

are analysed separately. However, the figure is approximately valid for low marker allele frequencies.

The numeric examples in Figure 4 indicate that when marker allele frequencies are low and the number of parental candidates is not excessive, pooling of progeny is advisable for determining the true two parents.

When contamination by pollen or seed has occurred, there is actually a progeny mixture, and a set of parents involved in its production. Certain modifications may then be required in the procedure. As above, a parental candidate combination may be eliminated whenever the progeny pool is missing *any* of its marker alleles. This strong elimination finally results in a narrow selection of parents, each having produced a large proportion of the progeny. With low allele frequencies, the power may be roughly approximated from expression (6). However, in order to compensate for the contamination, pool size must be increased. With 50% contamination, the adequate pool sizes corresponding to the above mentioned risk levels are N_p = 16, 21 and 29 (Table 1). Alternatively, a parental candidate combination may only be eliminated if the progeny pool is missing *both* marker alleles of either candidate. This more moderate elimination retains all the candidates which have produced at least one progeny individual in the pool.

Determining progeny by applying progeny candidate pools

A true progeny individual of a known parent pair $a_i/a_j \times a_k/a_l$ is to be determined from a set of N progeny candidates. The set is subdivided into pools of size N_p, and all pools which cannot contain a true progeny individual are eliminated. Regarding a single marker locus, the probability of elimination of a progeny candidate pool (r) is

$$P_e\{r\} = \frac{\left[\begin{matrix} N-n_{i+j} \\ N_p \end{matrix}\right] + \left[\begin{matrix} N-n_{k+l} \\ N_p \end{matrix}\right] - \left[\begin{matrix} N-n_{i+j+k+l} \\ N_p \end{matrix}\right]}{\left[\begin{matrix} N \\ N_p \end{matrix}\right]} , \tag{7}$$

and the relative power of elimination of a progeny candidate (t) with pooling is

$$RP_e\{t\} = \frac{N_p(N+2)}{(N+2N_p)} \cdot \frac{P_e\{r\}}{\left[1 - \dfrac{n_{i+j} + n_{k+l} - n_{i+j+k+l}}{N}\right]} . \tag{8}$$

The numeric examples show that in terms of the number of PCR reactions, the efficiency of determining the true progeny individuals can be substantially increased by applying progeny candidate pools (Figure 5).

Discussion

Ways of utilizing pooling in determining relatives were considered above for a diploid and cross-breeding plant species. After some modifications, self-pollination could also be allowed. Polyploidy would result in more complicated formulas and more stringent requirements *e.g.* for the informativeness of the marker type.

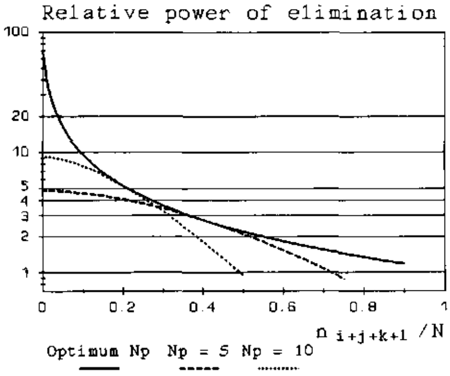


Figure 5. Relative power of elimination (RP_e) of progeny candidates per PCR reaction with applying progeny candidate pools compared to without applying pools. Without pooling $RP_e = 1$. Marker genotypes of the parents are a_i / a_j × a_k / a_l, and $n_{i+j+k+l} / N$ = proportion of candidates containing at least one parental allele. N = No. of progeny candidates ($N = 200$), and N_p = pool size. Conservative examples in which $n_{i+j+k+l} = n_{i+j} + n_{k+l}$

The probabilities of elimination (P_e) were given above for a single marker locus. Considering several unlinked marker loci (*j*), the combined probability of elimination is

$$P_e = 1 - \prod_j (1 - P_{e_j}) \,. \tag{9}$$

A multitude of DNA marker types are known at present. Since in the cases studied above, low marker allele frequencies appeared desirable, polyallelic markers, *e.g.*, microsatellites (Rafalski & Tingey 1993, Kauppinen *et al.* 1994) should preferably be used. A substantially smaller reduction in the number of PCR reactions is expected applying biallelic marker types, *e.g.*, RAPDs (Williams *et al.* 1993a) or AFLP markers (Zabeau & Vos 1993). However, savings are still achieved in the number of DNA extractions, which is of importance, particularly in plants (Åkerman *et al.* 1994). With AFLP markers, codominance can be utilized, whereas the information content of RAPDs is further reduced by dominance. In addition, missing bands sometimes occur with RAPD markers, perhaps due to their short primers. Therefore, in pooled-progeny analysis of parentage with RAPDs, the '+' results were only utilized (Åkerman *et al.* 1994). This precaution renders the informativeness of RAPD markers insufficient *e.g.* for the determination of the unknown parents of progeny mixtures applying a progeny pool.

Another possible way of reducing the number of PCR reactions is using marker types with a very large number of bands produced per lane, e.g., DAF (Caetano-Anollés *et al.* 1991), minisatellite repeat coding markers (Jeffreys *et al.* 1991), or even AFLPs with a very slight selection of bands. However, such marker types may not be as readily applied in the pooling procedures, because their banding patterns may be hard to interpret in pooled materials. Difficulties in distinguishing between certain alleles in a pool may sometimes also occur with *e.g.* RFLP markers and microsatellites. Especially with dinucleotide repeat microsatellites, the intense allelic band is often accompanied by a cluster of minor bands, possibly due to slippage during PCR amplification (Litt & Luty 1989, Schlötterer & Tautz 1992). In pooled materials, these extra minor bands may confuse scoring of adjacent allelic bands. A solution to this problem is combining such non-separable pair of alleles into a new, synthetic allele which can then be used in the procedures.

An additional application of pooling is the screening for contamination in the progeny of a known pair of parents. The occurrence of alien marker alleles due to illegitimate pollen or seed is followed in a set of pooled samples taken from the progeny. Alleles not present in the parents but common in the suspected source of contamination are preferably chosen, and the maximum pool size is applied. Much more progeny can be screened and thus a higher resolution is achieved by studying pools rather than the individuals separately.

The maximum pool size is restricted by the 'amplification potential' of the marker alleles, *i.e.*, their detection limit in bulked DNA (Williams *et al.* 1993b). Hence, depending on marker type and locus, the maximum pool size varies from about 10 with RFLPs (Michelmore *et al.* 1991), 10-100 with RAPDs (Michelmore *et al.* 1991, Williams *et al.* 1993b, Yu & Pauls 1993, Åkerman *et al.* 1994), 100 with solid-phase minisequencing markers (Syvänen *et al.* 1992), to 1000 with Blocker-PCR markers or when analysed using SSCP, even to 10,000 (Seyama *et al.* 1992). If necessary for not exceeding the maximum size, single pools can be subdivided, though at some loss of efficiency.

References

Åkerman, S., J. Tammisola, S.P. Lapinjoki, H. Söderlund, V. Kauppinen, A. Viherä-Aarnio, M. Regina & R. Hagqvist, 1994. RAPD markers in pooled-progeny analysis of parentage in European white birch (*Betula pendula* Roth). Can J Forest Res: submitted.

Caetano-Anollés, G., B.J. Bassam & P.M. Gresshoff, 1991. DNA amplification fingerprinting using very short arbitrary oligonucleotide primers. Bio/Technology 9: 553-557.

Ellstrand, N.C., 1984. Multiple paternity within the fruits of the wild radish, *Raphanus sativus*. Am Nat 123: 819-828.

Kauppinen, V., S. Lapinjoki, M. Regina, H. Söderlund, J. Tammisola, A. von Wright & S. Åkerman, 1994. Method for capturing and selection of microsatellite type DNA markers. Finnish pat. appl. No. 94 1765.

Litt, M. & J.A.Luty, 1989. A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. Am J Hum Genet 44: 397-401.

Michelmore, R.W., I. Paran & R.V. Kesseli, 1991. Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. Proc Natl Acad Sci USA 88: 9828-9832.

Rafalski, J.A. & S.V. Tingey, 1993. Genetic diagnostics in plant breeding: RAPDs, microsatellites and machines. Trends in Genetics 9: 275-280.

Schlötterer, C. & D. Tautz, 1992. Slippage synthesis of simple sequence DNA. Nucl Acid Res 20: 211-215.

Seyama, T., T. Ito, T. Hayashi, T. Mizuno, N. Nakamura & M. Akiyama, 1992. A novel blocker-PCR method for detection of rare mutant alleles in the presence of an excess amount of normal DNA. Nucl Acid Res 20: 2493-2496.

Syvänen, A.-C., E. Ikonen, T. Manninen, M. Bengtström, H. Söderlund, P. Aula & L. Peltonen, 1992. Convenient and quantitative determination of the frequency of a mutant allele using solid-phase minisequencing: application to aspartylglucosaminuria in Finland. Genomics 12: 590-595.

Taylor, B.A., A. Navin & S.J. Phillips, 1994. PCR-amplification of simple sequence repeat variants from pooled DNA samples for rapidly mapping new mutations of the mouse. Genomics 21: 626-632.

Williams, J.G.K., M.K. Hanafey, J.A. Rafalski & S.V. Tingey, 1993a. Genetic analysis using RAPD markers. Meth Enzymol 218: 704-740.

Williams, J.G.K., R.S. Reiter, R.M. Young & P.A. Scolnik, 1993b. Genetic mapping of mutations using phenotypic pools and mapped RAPD markers. Nucl Acid Res 21: 2697-2702.

Yu, K. & K.P. Pauls, 1993. Rapid estimation of genetic relatedness among heterogeneous populations of alfalfa by random amplification of bulked genomic DNA samples. Theor Appl Genet 86: 788-794.

Zabeau, M. & P. Vos, 1993. Selective restriction fragment amplification: a general method for DNA fingerprinting. Eur Pat Off EP0534858A1, 43 p.

Comparison of different approaches to interval mapping of quantitative trait loci

H.F. Utz & A.E. Melchinger, Institut für Pflanzenzüchtung, Saatgutforschung und Populationsgenetik der Universität Hohenheim, D-70593 Stuttgart, Germany

Key words
molecular markers, QTL mapping, simulation

Summary

By using simulation, we compared three methods for interval mapping of quantitative trait loci (QTLs): the conventional standard procedure of Lander & Botstein and two new methods using linked and/or unlinked markers as cofactors in a multiple regression approach. We assumed a genome of 10 "chromosomes", 200 cM long, covered with equidistant (20 cM) markers and 14 QTLs of different size and sign.

The power of detecting QTLs decreased and the bias of estimated QTL effects increased for lower heritabilities and smaller sample sizes. Methods using cofactors showed a larger power of detecting QTLs and reduced bias and sampling error of estimated QTL positions and effects than the standard method without cofactors. Therefore, we recommend the use of selected cofactors in QTL mapping, especially when the possibility of several QTLs on the same chromosome cannot be excluded.

Introduction

Mapping of genes affecting quantitative traits receives growing attention in breeding. It is anticipated that more efficient selection strategies can be applied, once the positions and effects of important quantitative trait loci (QTL) are known. Currently, the interval mapping method of Lander & Botstein (1989) is routinely used for mapping of QTLs as implemented in the program MAPMAKER/QTL of Lincoln *et al.* (1993). With this method, each chromosome is scanned for the most likely position of QTLs in each marker interval.

Problems with this method arise especially when several QTLs are located on the same chromosome, because estimates of the position and the effect of a QTL can be biased by adjacent QTLs (*cf.* Stam 1991, Martinez & Curnow 1992, Haley & Knott 1992, Van Ooijen 1992). Two closely linked QTLs are often identified as a single ("ghost") QTL located at an intermediate position. QTLs with effects of opposite sign may cancel each other so that they cannot be detected. Furthermore, precision of QTL estimates is reduced, because genotype frequencies of several QTLs are generally non-orthogonal in populations of manageable size. Therefore, simply fitting individual QTLs is inadequate and several QTLs should be fitted simultaneously. When working with MAPMAKER/QTL, a simultaneous search for multiple QTLs is performed only in rare cases that are selected on a rather subjective basis related to increased computational requirements and difficulties concerning parameter estimation and model identifiability.

A systematic search in two dimensions for two linked QTLs was devised by Haley & Knott (1992) using a regression procedure. However, this strategy becomes computationally rather cumbersome if more than two QTLs are fitted simultaneously. As markers flanking a QTL absorb most of the variation caused by that QTL, an alternative approach is using interval mapping with other markers as cofactors (Jansen 1993, Jansen & Stam 1994, Zeng 1994).

These newly suggested mapping procedures differ in certain details such as the treatment of missing values or the selection of important cofactors. For practical applications in biology and breeding, this raises the question whether these methods lead to different results and which method should be preferred in a given situation.

The objectives of this paper are to compare different procedures for interval mapping with regard to (1) the power of detecting QTLs and (2) the bias and the sampling error of estimates of the positions and effects of QTLs. This was done using simulations with different population sizes, heritabilities and assumptions about locations and effects of QTLs. All computations were based on a multiple regression approach. The influence of the generation on QTL estimates was investigated by comparing results of F_2 and F_3 populations as well as by including results of parents.

Methods

The Model

We assumed a completely additive genetic model which applies, for example, to testcross progenies in the absence of epistasis (Cowen 1988). Accordingly, the genotypic means of genotypes QQ, Qq and qq at a QTL are $m+a$, m and $m-a$, respectively. In order to arrive at an interval procedure which combines interval mapping with regression analysis on flanking markers of other QTLs, Zeng (1994) used the following model

$$y_j = m + b^* x^*_{jl} + \sum_k b_k x_{jk} + \epsilon_j \,. \tag{1}$$

In (1), y_j denotes the value of the trait considered for the j^{th} individual, m is the grand mean, b^* is the effect of the putative QTL in marker interval $(l, l+1)$, x^*_{jl} is a random variable, which assumes values 1, 0 or -1 with probabilities depending on the genotypes of markers l and $l+1$ and the position of the putative QTL, b_k is a regression coefficient related to the k^{th} cofactor, x_{jk} is a dummy variable or cofactor, which assumes values 1, 0 or -1 depending on whether the genotype of individual j at marker locus k is $\mathrm{M}_k\mathrm{M}_k$, $\mathrm{M}_k\mathrm{m}_k$ or $\mathrm{m}_k\mathrm{m}_k$, respectively, and ϵ_j is an independent residual.

Composite Interval Mapping

Zeng (1994) showed that maximum likelihood (ML) estimates for the parameters in the above mixture model situation can be obtained via the expectation/conditional maximization (ECM) algorithm. However, the necessary computations are too time-consuming for the present simulation study. For this reason, we extended the regression approach of Haley & Knott (1992) and Martinez & Curnow (1992) to include other (linked and unlinked) markers as cofactors. Thus, we used equation (1) but instead of x^*_{jl}, we used as regressor the conditional expectation of x^*_{jl} given the observed genotypes at the flanking marker loci l and $l+1$. Conditional expectations for all possible flanking marker genotypes in an F_2 population were calculated according to the formulae in Table 1 of Haley & Knott (1992). Investigations by these authors and our own results (not shown) demonstrated that regression and ML yield almost identical LOD profiles and parameter estimates.

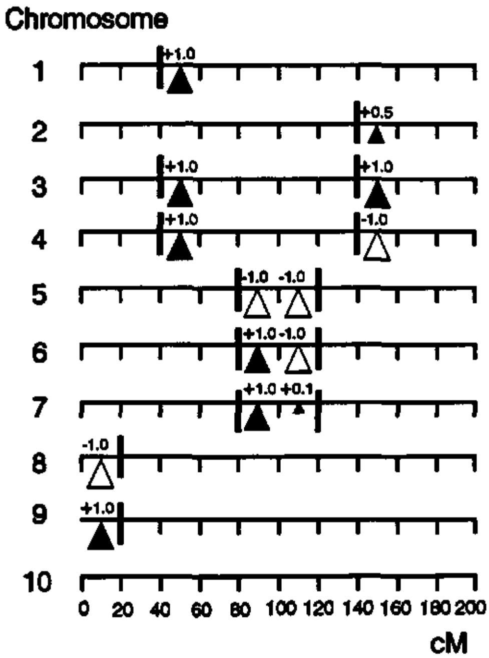
In the approach outlined above, phenotypic and genotypic data of parental lines and their F_1 progeny can be included without any further modifications. Furthermore, the approach can also be applied to F_3 populations using the relevant formulae for conditional expectations presented by Dillmann & Melchinger (1994).

Cofactors x_{jk} may comprise all markers or a selected subset adjacent to putative QTLs. Adopting the notation of Zeng (1994), we considered three procedures (using as cofactors flanking markers of QTLs, as shown in Figure 1):

Method I: Composite interval mapping using linked and unlinked cofactors but excluding those cofactors flanking the QTL to be fitted;

Method II: Semi-composite interval mapping using only unlinked cofactors;

Figure 1. Genetic map with 14 QTL (Δ) on 10 chromosomes (200 cM long), each with 11 markers (except for chr. 9) spaced 20 cM apart. Markers used as cofactors are indicated by a bold vertical bar. Numbers and symbols refer to the size and sign of QTL effects *a*



Method III: Conventional interval mapping without using any cofactors.

In the search for significant QTLs, Jansen (1993) proposed a multi-stage decision procedure, whereas Zeng (1994) advocated a simple interval (likelihood ratio) test. Zeng's method compares the maximized likelihood (L_1) of the model including the putative QTL with the maximized likelihood (L_0) of the model without the QTL; in both models all selected cofactors are included. Since this procedure is similar to the strategy of the program MAPMAKER/QTL and less complicated for the user, we employed the test procedure suggested by Zeng (1994). The threshold of the LOD score ($\log_{10}(L_1/L_0)$) for a putative QTL to be significant was set equal to 2.5 for all three methods. This corresponds to a 30% critical value for an overall test with 99 intervals based on a chi-squared distribution with 2 df (1 df for the position and 1 df for the effect). Power was calculated as the proportion of the simulations in which the QTL was detected within the 20 cM interval containing the QTL. As is common practice, estimates of QTL positions and QTL effects were obtained at the maximum of the LOD score curve in the relevant region.

Simulations

Simulations were performed assuming a genome of 10 "chromosomes", 200 cM long, each covered with 10 or 11 markers separated by 20 cM intervals (Figure 1). The trait is affected by 14 QTL with positions and effects given in Figure 1. In addition, we assumed normally distributed "noise" variables ϵ with common variance σ_ϵ^2 . The values of the heritability of the trait were set equal to 0.4, 0.7, and 1.0. The marker map and cofactors (one per QTL) were treated as being known for each replicate. Altogether, we generated a basic sample of 16500 F_2 or F_3 individuals, which was partitioned into 165 samples of size $n = 100$, 55 samples of size $n = 300$ or 27 samples of size $n = 600$.

Results

For chromosome 4 (2 QTL of equal size with opposite signs), Figure 2 shows typical LOD profiles obtained with the three methods using one of the samples. All three methods yielded a clear peak for the LOD profile in the vicinity of the QTL, exceeding 2.5 except for one QTL with Method III. Maximum LOD scores were largest for Method I, slightly smaller for Method II, and smallest for Method III.

These characteristics are also reflected in the summary statistics for the simulation results concerning "chromosome" 1 (Table 1). The power of detecting a QTL with $a = 1.0$ (explaining 8.1% of the genetic variance in an F_2 population) was generally largest for Method I, almost as large for Method II, but substantially smaller for Method III, irrespective of the sample size and the heritability of the trait. The power was fairly small for $h^2 = 0.4$ with sample size $n = 300$, but increased considerably for $h^2 = 0.7$ and $h^2 = 1.0$, even with small sample sizes. For all three methods the estimated positions of the QTL on chromosome 1 were close to the true value (50 cM). With respect to the

Figure 2. LOD profile of Methods I, II, and III for one sample of simulation with two QTL at 50 and 150 cM with effect $a = 1.0$ and $a = -1.0$, respectively (sample size $n = 300$, $h^2 = 0.7$)

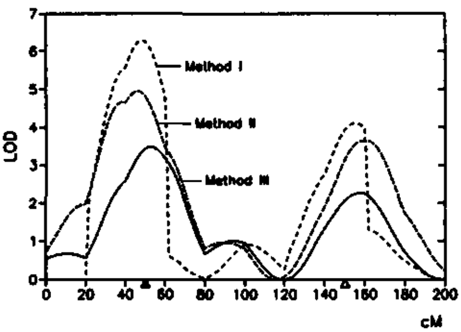


Table 1. Summary statistics of estimates of positions and effects of QTL on chromosome 1 (1 QTL at position 50 cM with additive effect a = 1.0) from r samples of simulations with population size n and heritability h^2 of the trait, using three different methods of QTL analysis

Heritability:	0.4			0.7			1.0		
Method:	I	II	III	I	II	III	I	II	III
n = 100, r = 165:									
Power[†] (%)	6.1	6.1	4.2	30.9	33.9	11.5	71.5	65.6	19.4
Position[‡]	47.4	52.1	49.3	47.4	50.2	47.7	48.3	49.6	49.8
	6.3	*17.9*	*27.0*	*15.2*	*17.9*	*16.2*	*7.9*	*11.6*	*17.3*
Effect[‡]	1.97	1.96	1.84	1.36	1.41	1.77	1.09	1.09	1.51
	0.22	*0.21*	*1.03*	*0.45*	*0.24*	*0.21*	*0.21*	*0.21*	*0.21*
n = 300, r = 55:									
Power	40.0	34.5	23.6	90.9	81.8	56.4	92.7	92.7	72.7
Position	48.3	47.3	50.6	49.7	49.9	50.9	49.1	49.4	49.1
	8.3	*14.4*	*13.5*	*6.6*	*8.3*	*14.3*	*5.4*	*5.5*	*12.5*
Effect	1.20	1.20	1.33	1.05	1.05	1.15	1.02	1.01	1.07
	0.20	*0.19*	*0.17*	*0.17*	*0.16*	*0.19*	*0.14*	*0.12*	*0.19*
n = 600, r = 27:									
Power	74.1	70.4	55.6	96.3	96.3	81.5	100.0	100.0	96.3
Position	47.2	47.8	48.8	50.4	50.8	50.4	49.9	49.9	49.2
	7.3	*8.6*	*10.5*	*3.8*	*3.8*	*9.3*	*2.2*	*2.2*	*5.8*
Effect	1.02	1.03	1.07	1.03	1.03	1.04	1.02	1.02	1.02
	0.19	*0.18*	*0.17*	*0.15*	*0.15*	*0.19*	*0.10*	*0.10*	*0.16*

[†] Empirical estimate of the power of the test for a QTL in the interval (40 cM, 60 cM).

[‡] Mean and standard deviation (in italics) estimates based on the samples yielding a significant QTL

standard error of position estimates, Method I was clearly superior to the other two methods for low heritabilities ($h^2 = 0.4$) or small sample sizes ($n = 100$) and Method III was clearly inferior for high heritabilities ($h^2 \geq 0.7$). For all three methods estimates of QTL effects a were seriously biased (up to 97%) for low heritabilities ($h^2 = 0.4$, 0.7) and small sample sizes ($n = 100$, 300). The bias was consistently largest for Method III and it could only be ignored for Methods I and II with $n = 600$ or $h^2 = 1.0$. Standard deviations of estimated QTL effects were similar for all three methods, heritabilities, and population sizes. Occasional deviations from the general trends for estimated positions and effects of QTLs were attributable to extreme values in some samples.

Table 2 summarizes the simulation results for other chromosomes for sample size $n = 300$ and heritability $h^2 = 0.7$. Compared to the QTL on chromosome 1 with $a = 1.0$, the power of detecting the QTL on chromosome 2 with $a = 0.5$ (corresponding to 2.0% of the genetic variance) was about one third. Estimates of QTL position were still unbiased but had slightly smaller precision. In contrast, estimates of QTL effects were on average grossly inflated (50%) for Methods I and II and even more (> 100%) for Method III.

Chromosomes 3 and 4 (2 linked QTL with a map distance 100 cM, $|a| = 1.0$, same

Table 2. Summary statistics of estimates of positions and effects of QTL on chromosomes 2, 3, 4, and 7 from r = 55 samples of simulations with sample size $n = 300$ and $h^2 = 0.7$ for three different methods of QTL analysis

Chromosome; QTL position and effect	Method		
	I	II	III
2; 150 cM, a = 0.5:			
Power[†] (%)	30.9	34.5	9.1
Position[‡]	147.6 ± *9.3*	151.6 ± *18.8*	153.2 ± *10.5*
Effect[‡]	0.75 ± *0.12*	0.75 ± *0.10*	1.03 ± *0.12*
3; 50 cM, a = 1.0 [QTL2: 150 cM, a = 1.0]:			
Power	98.2	69.1	67.3
Position	49.1 ± *6.0*	55.8 ± *10.8*	56.1 ± *11.0*
Effect	1.02 ± *0.16*	1.20 ± *0.18*	1.22 ± *0.21*
4; 50 cM, a = 1.0 [QTL2: 150 cM, a = −1.0]:			
Power	100.0	76.4	30.9
Position	49.4 ± *5.8*	46.5 ± *7.9*	45.0 ± *10.7*
Effect	1.02 ± *0.14*	0.88 ± *0.13*	1.03 ± *0.13*
7; 90 cM, a = 1.0 [QTL2: 110 cM, a = 0.1]:			
Power	69.1	80.0	61.8
Position	94.6 ± *12.6*	91.9 ± *17.0*	92.4 ± *19.2*
Effect	1.06 ± *0.33*	1.07 ± *0.20*	1.15 ± *0.23*

[†] Empirical estimate of the power of the test for a QTL in the interval spanned by flanking markers.

[‡] Mean and standard deviation (in italics) of estimates based on the samples which are significant in the test (LOD > 2.5) in the relevant interval

and opposite signs of QTL effects, respectively) showed similar results with Method I (Table 2): the power of detecting a QTL was almost 100% and QTL positions and QTL effects were estimated without bias. In contrast, the power for detecting a QTL was reduced about 25% with Method II for both chromosomes. Estimates for QTL position were biased, *i.e.*, estimated distances between QTLs were too small for linked QTLs in coupling phase, and too large for linked QTLs in repulsion phase. Furthermore, estimates of QTL effects were inflated and deflated, respectively. Method III yielded similar results as Method II for chromosome 3, but had a small power for detecting a QTL for linked QTLs in repulsion phase. These trends are stronger for the tightly linked QTLs (map distance 20 cM) on chromosomes 5 and 6 (data not shown). For chromosome 6, the power of detecting a QTL with Methods II and III (no linked cofactors in the model) dropped even to 0%.

Finally, chromosome 7 represents the case of a major QTL ($a = 1.0$) in tight coupling phase linkage (map distance 20 cM) with a minor QTL ($a = 0.1$). The estimated position of the major QTL showed only negligible bias towards the position of the minor QTL but its standard deviation was for all three methods considerably larger than the corresponding values for the comparable QTL on chromosome 1.

Discussion

The results from our simulations clearly demonstrate that use of selected cofactors in interval mapping (Methods I and II) is superior to the conventional standard procedure (Method III). Substantial improvement in the power of detecting a QTL, reduction of bias and precision of estimated QTL positions and QTL effects were found consistently. Zeng (1994) and Jansen & Stam (1994) also reported a larger power and larger values of likelihood ratio statistics with Method II (comparison of Model A_2 with B_2 in Jansen & Stam) indicating that most of the genetic variation due to segregating QTLs on other chromosomes can be effectively removed by the use of cofactors.

Zeng (1994) found a considerably smaller power for Method I in comparison with Method II. He emphasized that Method I is an interval test, *i.e.*, a test for the presence of a QTL in the interval considered rather than a test for the presence of a QTL on the entire chromosome, as is the case for Methods II and III. In contrast to his findings, we found a similar or larger power for Method I in comparison with Method II except for chromosome 7. This discrepancy may have two reasons: (1) Zeng (1994) included all (linked and unlinked) markers as cofactors, whereas we employed only one cofactor adjacent to known QTLs; (2) closely linked non-informative cofactors reduce the power to a greater extent than loosely linked or unlinked non-informative cofactors as follows from the derivations given by Zeng (1993). In our study, we only had a fairly non-informative closely linked cofactor on chromosome 7, and this caused a smaller power for Method I compared with Method II, as expected.

Choosing the most influential markers as cofactors is critical for Methods I and II because the use of non-informative cofactors reduces the power of detecting a QTL. In our study, the choice of cofactors did not depend on the observations but was fixed, based on the position of known QTLs. Therefore, Methods I and II may be less superior to Method III in real situations, when the locations of QTLs are unknown. Jansen (1993) described a rather sophisticated procedure for selecting the set of markers to be used as cofactors, starting with multiple regression and the method of backward elimination. He proposed using Akaike's information criterion as a tool for model selection. Jansen & Stam (1994) recommended that the number of cofactors should not exceed 2 √(number of observations). Zeng (1994) suggested to select a few markers on each chromosome by stepwise regression. In order to warrant unbiasedness of estimates of QTL position and QTL effects, Zeng (1993) advocated to include as additional cofactors the markers on the right and left-hand side of the interval to be tested. However, as pointed out above, this may result in a substantial reduction of the power when these cofactors are non-informative.

Our simulations corroborate the conclusion of Zeng (1994) that separation of tightly linked (< 20 cM) QTLs located in adjacent marker intervals remains still an unsolved problem. Haley & Knott (1992) also reported that QTLs with a distance of 20 cM (as was the case for chromosomes 5 and 6) could hardly be separated.

A smaller power of detecting a QTL was generally associated with a larger bias of the estimated QTL effects a. An exception was observed for chromosome 4 with Method III: the upward bias caused by the smaller power was compensated by a downward bias caused by the second QTL in repulsion phase on the same chromosome (100 cM distant). In conventional interval mapping (Method III), bias of the QTL effects precludes a realistic judgement about the prospects of marker-assisted selection because it parallels overestimation of the phenotypic variance explained by the putative QTL. It is noteworthy that with Method III and a given simulation sample, MAPMAKER/QTL always yielded larger R^2 values than our program based on the regression approach.

QTL mapping based on F_3 instead of F_2 generations (results not shown) indicated two trends: the power of detecting a QTL was slightly reduced in the F_3, and correspondingly, the bias of the estimated QTL effects was increased. This is very likely the result of increased recombination between adjacent markers. Use of cofactors in Methods I and II is even more superior to Method III with F_3 populations, as the "background genetic" noise is increased due to a larger genetic variation among F_3 individuals. Further research is warranted to compare F_2 with F_3 populations and later generations obtained by inbreeding with regard to the power of QTL detection and precision of QTL estimates obtained with Method I for different marker densities, heritabilities and other factors of importance.

Jansen & Stam (1994) suggested that use of cofactors in QTL mapping make it possible to include the parents and the F_1 progeny in addition to F_2 progeny. Preliminary simulations showed that with the latter procedure the LOD score and the power are increased if the difference in the parents has the same sign as the difference for the two QTL genotypes present in the parents. In the opposite case, the LOD score is reduced. For many applications in plant breeding, *e.g.*, marker-assisted selection, it is of special interest to detect unfavourable alleles in the superior parent, because combining all favourable alleles in a single genotype should lead to a transgression of the better parent. Thus, including the parents will not necessarily contribute to an increased efficiency and precision of QTL mapping and cannot be recommended as a general procedure.

Acknowledgements

We are indebted to Dr. G. Seitz for providing his program to generate the basic sample used in our simulations. We also owe gratitude to C.S. Haley and J.W. van Ooijen, who provided their computer programs for making numerical comparisons.

References

Cowen, N.M., 1988. The use of replicated progenies in marker-based mapping of QTLs. Theor. Appl. Genet. 75: 857-862.

Dillmann, C. & A.E. Melchinger, 1994. Three-locus genotype frequencies in selfed generations derived from F_1 crosses with applications to QTL mapping. (In preparation).

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

Jansen, R.C., 1993. Interval mapping of multiple quantitative trait loci. Genetics 135: 205-211.

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

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

Lincoln, S.E., M.J. Daly, & E.S. Lander, 1993. Mapping genes controlling quantitative traits using MAPMAKER/QTL version 1.1: A tutorial and reference manual. Whitehead Institute for Biomedical Research, Cambridge, MA.

Martinez, O. & R.N. Curnow, 1992. Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. Theor. Appl. Genet. 85: 480-488.

Stam, P. 1991. Some aspects of QTL analysis. *In*: Proc. 8th Meeting Eucarpia Section Biometrics in Plant Breeding, Research Institute of Agroecology and Soil Management, Hrusovany near Brno, 23-32.

Van Ooijen, J.W., 1992. Accuracy of mapping quantitative trait loci in autogamous species. Theor. Appl. Genet. 84: 803-811.

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

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

Comparison of a single-QTL model with an approximate multiple-QTL model for QTL mapping

J.W. van Ooijen, DLO-Centre for Plant Breeding and Reproduction Research, P.O. Box 16, 6700 AA Wageningen, The Netherlands

Summary

The essentials of several models used for QTL mapping are described. A small computer simulation study was carried out to investigate the gain of using an approximate multiple-QTL model ('MQM mapping', Jansen 1993, 1994) over a single-QTL model ('interval mapping', Lander & Botstein 1989) for the case of several unlinked QTLs. It appeared that both the power to detect a QTL and the precision of localizing a QTL can be enhanced considerably.

Introduction

The advent of molecular markers has greatly facilitated the genetic analysis of quantitative traits. In general such an analysis consists of detecting and localizing the genes affecting a trait and of estimating the effects on that trait that are associated with the alleles. We have seen an important development of statistical methods that can be used in such an analysis. Hereafter we will consider the essentials of some of these methods.

The principle of any statistical QTL mapping method is the detection of a genetic effect masked by a certain amount of residual variation. The traditional approach to such a problem would be to perform an analysis of variance (ANOVA) based on QTL genotype classes. However, the individuals of a segregating progeny cannot be classified according to the QTL genotype classes because these are disguised by residual variation. The solution to this problem is, of course, the employment of markers. When a marker is tightly linked to a QTL, the classification into marker genotypes will correspond closely to the classification according to the (unknown) QTL genotypes. The best statistical approach would therefore be performing ANOVAs based on marker genotype classes.

Problems arise when the linkage of the marker to the QTL is not close. Marker genotype classes supposed to consist of a single QTL genotype, will be 'contaminated' with other QTL genotypes. As a consequence, the differences between the marker classes will reduce with increasing distance of marker to QTL, resulting in a loss of power for QTL detection. Apart from this the error terms in the model will not be identically distributed, thereby violating an ANOVA assumption. The solution is the use of a mixture model. Individual i with marker genotype m has a probability π_{mq} to be of QTL genotype q, which depends on the map distance of the marker to the QTL. In the mixture model the probability density of the quantitative trait value y_i of individual i is given by the so-called mixture density:

$$f(y_i) = \sum\nolimits_q \pi_{mq} f_q(y_i),$$

where the summation is over all QTL genotypes, and where $f_q(.)$ denotes the density corresponding to QTL genotype q. The parameters of the mixture model can be estimated by maximum likelihood, which can be done conveniently with an EM-algorithm. The mixture model is tested against a model with equal densities for all QTL genotypes, *i.e.*, a model with a zero genetic effect. The analysis may be considered as a weighted analysis of variance. When simultaneously two neighbouring markers are used to calculate the genotype probabilities for a QTL in between, it becomes the method introduced by Lander & Botstein (1989) known as 'interval mapping'.

So far, the model presented assumes the segregation of a single QTL in a certain amount of residual variation (single-QTL model). When several QTLs are segregating, however, the residual variation consists of an environmental as well as a genetic component. Lander & Botstein (1989) already recognized that the power of the analysis could be enhanced by fitting a multiple-QTL model, due to the reduction of the residual variance. In addition, when there are two or more linked (and also segregating) QTLs a single-QTL model may even map a QTL at the wrong position (Haley & Knott 1992, Martinez & Curnow 1992). The mixture density for a multiple-QTL model is also:

$$f(y_i) = \sum\nolimits_q \pi_{mq} f_q(y_i),$$

but here q refers to the *multiple*-QTL genotype, *i.e.*, the genotype of all QTLs in the model, and m to the genotype of all markers neighbouring these QTLs. It should be noted, that the number of components in the mixture density increases exponentially with the number of modelled QTLs. As a consequence the computations involved in solving such a model are unfeasible when the number of QTLs is large (Jansen 1993, Zeng 1993). To solve this problem, Jansen (1992, 1993), Jansen & Stam (1994) and Zeng

(1993, 1994) introduced approximate multiple-QTL models that combine interval mapping of a single QTL with multiple linear regression to account for other QTLs.

In the approximate multiple-QTL model of Jansen (1993), the so-called 'MQM mapping' method (Jansen 1994), markers are selected and subsequently used as cofactors in an interval mapping procedure for mapping a single QTL. It is assumed that the selected marker cofactors absorb the genetic effects of closely linked QTLs, thereby enhancing the power to detect other QTLs. The mixture density reads:

$$f(y_i) = \sum_q \pi_{mq} f_{Cq}(y_i),$$

where $f_{Cq}(\cdot)$ denotes the density corresponding to both the (single) QTL genotype q and (multiple) cofactor markers genotype C, with a mean μ_{Cq}, which depends on cofactor markers C, for instance for an autogamous species:

$$\mu_{Cq} = \mu_q + \sum_c (g_c a_c + h_c d_c),$$

where μ_q is the overall mean of (single) QTL genotype q, a_c and d_c are the additive and dominance effect associated with marker cofactor c, and g_c (= 1, 0, or -1) and h_c (= 0, 1, or 0) are indicator variables for the genotype of cofactor marker c (= CC, Cc, or cc, respectively). The model is tested against a model with equal densities for the QTL genotypes, *i.e.*, the μ_qs are equal. In comparison with the multiple-QTL model this model is approximate, first, because it assumes complete linkage between a cofactor marker and the QTL, of which it is supposed to absorb the effect, and secondly, because the model neglects gene interaction effects.

The gain of an (approximate) multiple-QTL model over a single-QTL model in the case of two or more *linked* QTLs was obvious after the illustrations of such situations by Haley & Knott (1992) and Martinez & Curnow (1992). To investigate the gain of using a multiple-QTL model over a single-QTL model for the case of several *unlinked* QTLs a small simulation study was carried out.

Computer simulation study

Ten F_2 populations of 200 plants each were simulated. The genome consisted of twelve chromosomes of 120 centiMorgan each with a segregating marker at every five centiMorgan. This specific configuration was chosen to be able to relate this study to a previous more extensive simulation study (Van Ooijen 1992). Six chromosomes carried a segregating QTL at the 62.5 cM position, in between the 13th and 14th marker. Three of these QTLs were given an additive genetic effect (dominance was absent) with which they were expected to explain 10% of the total (= genetic + residual) variance (10%-QTL), and three other QTLs were given an effect of 5% expected explained variance (5%-QTL). The remaining six chromosomes were without QTL.

Each population was analysed both with interval mapping (Lander & Botstein 1989) and with MQM mapping (Jansen 1993, 1994). In practice the cofactors in the approximate multiple-QTL model have to be selected, in some way, based on the experimental data, whereas in this simulation study the most informative markers were used: the markers closest to the QTL were used as cofactors (map distance marker to QTL was 2.5 cM). Consequently, this simulation study indicates more or less the maximum attainable resolution of MQM mapping. A marker cofactor was not contained in the model when a QTL was fitted on the same chromosome, thus the model contained five cofactors when fitting a QTL on the chromosomes containing a QTL, whereas it contained six for the other chromosomes.

The effect of the map distance of the marker cofactors to the QTLs was investigated by analysing the data also with different markers as cofactors. First, the markers with a map distance of 2.5 cM to the QTL were replaced by markers at 7.5, 17.5, 27.5, or 37.5 cM. Second, two markers simultaneously at either side of the QTL (a so-called marker bracket) were taken as cofactors, at distances to the QTL given above.

The significance threshold for the detection of a QTL was taken as 3.7 LOD for both methods (Van Ooijen 1992). When a QTL was detected, the distance of the estimated map position to the real position of the QTL was determined and a 2-LOD support interval was constructed as described by Van Ooijen (1992). The length of this interval was determined.

Table 1. The number of detected QTLs for the ten simulated F_2s

Number of detected QTLs			Number of F_2s	
10%-QTL	5%-QTL		Single-QTL model	Approx. multiple-QTL model
2	0		1	-
2	2		1	-
3	0		4	2
3	1		4	4
3	2		-	2
3	3		-	2
		Total:	10	10

Table 2. The average length of the support intervals. The length is given in centiMorgan. The number on which an average is based, is given between brackets

Single-QTL model		Approx. multiple-QTL model	
10%-QTL	5%-QTL	10%-QTL	5%-QTL
33.4 (28)	27.8 (6)	23.2 (30)	32.3 (14)
Based upon QTLs detected by both methods:			
33.4 (28)	30.9 (5)	23.3 (28)	23.8 (5)

Results and discussion

Ten F_2 populations were simulated, each segregating for three 10%-QTLs and three 5%-QTLs. Interval mapping detected 28 of the 30 segregating 10%-QTL and 6 of the 30 5%-QTLs, whereas MQM mapping detected 30 and 14 of these QTLs, respectively. Neither method produced a false positive. Table 1 presents these results in more detail. The results for the single-QTL model are comparable to those of the extensive simulations presented by Van Ooijen (1992). It is evident that the approximate multiple-QTL mapping method has a greater power to detect QTLs. Especially the 5%-QTLs, which have a small chance of detection with a population size of 200, have a better chance of being detected when other segregating QTLs are taken into account. The 10%-QTLs already have a large chance of detection with interval mapping, so the gain with MQM mapping cannot be very large.

When a QTL was detected, a 2-LOD support interval was constructed. Such an interval is supposed to act as a 95% confidence interval, and as such it shows with its length the precision with which the detected QTL has been localized. The average lengths of the support intervals are presented in Table 2. The results for interval mapping are comparable to those in the paper by Van Ooijen (1992), although the average length for the 5%-QTLs was somewhat short. Presumably, this is due to the large variation of the support interval length, combined with the small number of detected 5%-QTLs (Van Ooijen 1992). With the multiple-QTL model the average length for the 10%-QTLs was reduced, but the opposite happened for the 5%-QTLs. This latter fact must be due to the larger number of detected 5%-QTLs the MQM mapping average is based upon: QTLs that remain undetected with interval mapping do get detected with MQM mapping, albeit relatively inaccurate. When we look at the averages based upon only those QTLs detected by both interval and MQM mapping, we see that for the 10%-QTLs as well as the 5%-QTLs the length of the support interval was reduced considerably (Table 2).

For detected QTLs the distance of the estimated map position to the simulated map position was determined. The averages are given in Table 3. When only the QTLs detected by both interval and MQM mapping are considered, the distance was smaller

Table 3. The average distance of the estimated to the simulated QTL position. Distances are given in centiMorgan. The number on which an average is based, is given between brackets

Single-QTL model		Approx. multiple-QTL model	
10%-QTL	5%-QTL	10%-QTL	5%-QTL
9.1 (28)	5.1 (6)	3.5 (30)	4.1 (14)
Based upon QTLs detected by both methods:			
9.1 (28)	5.5 (5)	3.5 (28)	5.5 (5)

with MQM mapping for 10%-QTLs whereas it was equal for 5%-QTLs. We don't see an enhancement for the 5%-QTLs with MQM mapping. This may be due to the small number the average is based upon.

Table 4 summarizes the maximum values that were obtained over the simulation runs. Here too, we observe a reduction in the length of the support interval for both the 5%-QTLs and the 10%-QTLs with the use of the approximate multiple-QTL model. The same holds for the distance of the estimated to the simulated QTL map position of the 10%-QTLs, whereas there was no reduction for the 5%-QTLs, presumably due to the small number the average is based upon.

In practice the cofactor markers in the approximate multiple-QTL model have to be selected in some way based on the experimental data. When interval mapping would be used for this purpose, the marker closest to the estimated map position of the QTL, or even two markers at either side of this position, could be used as cofactor(s). However, the estimated QTL position may sometimes be quite some distance away from the real position (Table 4). Therefore, the effect of the map distance of the cofactor markers to the QTLs on MQM mapping was investigated by analysing the data also with cofactor markers and marker brackets at several distances. The results for the 10%-QTLs are depicted in Figure 1, those for the 5%-QTLs were similar (data not shown). At a greater cofactor marker to QTL distance the support interval length was larger, although even at

Table 4. The maxima of the support interval length and the distance of estimated to simulated QTL position obtained in the simulation runs, based upon only the QTLs detected by both methods (28 10%-QTLs, 5 5%-QTLs). The lengths and distances are given in centiMorgan

Single-QTL model		Approx. multiple-QTL model	
10%-QTL	5%-QTL	10%-QTL	5%-QTL
Length of support interval:			
86	46	42	33
Distance of estimated to simulated QTL position:			
44	15	13	15

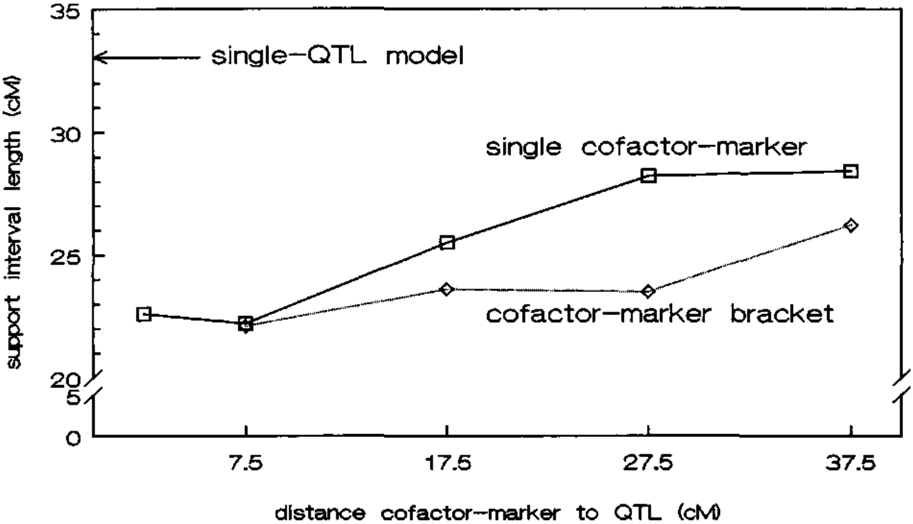


Figure 1. The effect of the map distance of the cofactor-marker to the QTL in the approximate multiple-QTL model. All points are based upon the same 26 10%-QTLs that were detected in all situations. The interval length obtained for the same cases with the single-QTL model is indicated at the y-axis

37.5 cM the results are still better than at the single-QTL model. Due to an increased probability of recombination, the power of the cofactor marker to absorb the effect of the linked QTL at a larger distance becomes smaller, and consequently, the mapping precision is reduced. The use of a cofactor marker bracket instead of a single marker was less affected by the cofactor marker to QTL distance. The reason for this must be the following: even when there is a single recombination between one of the cofactor markers in the bracket, the QTL effect can still be partly absorbed by the other marker.

Conclusion

This small simulation study showed that the power to detect a QTL and the precision of localizing a QTL can be enhanced considerably by using an approximate multiple-QTL model, such as MQM mapping, instead of a single-QTL model. The rationale behind it is the reduction of the residual variance by the use of cofactor markers, which absorb the genetic effects of closely linked QTLs. Of course, a gain is only possible, when there are more than one segregating QTLs. Also, the gain is expected to be modest when the QTLs have a relatively small genetic effect, because of two reasons. First, a QTL with a small effect will be mapped inaccurately, and therefore, the distance to the linked cofactor marker will, on average, be large. Secondly, even a complete absorption of the QTL effect would lead to just a modest reduction of the residual variance. In the simulation study six QTLs with a substantial effect were segregating: the variance caused by the QTL in relation to the unexplained residual variance, when all other QTLs are taken into account in the model, was 10/(10+55)=15% for the 10%-QTL and 5/(5+55)=8% for the 5%-QTL.

References

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

Jansen, R.C., 1992. A general mixture model for mapping quantitative trait loci by using molecular markers. Theor. Appl. Genet. 85: 252-260.

Jansen, R.C., 1993. Interval mapping of multiple quantitative trait loci. Genetics 135: 205-211.

Jansen, R.C., 1994. Controlling the type I and type II errors in mapping quantitative trait loci. Genetics: in press.

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

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

Martinez, O. & R.N. Curnow, 1992. Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. Theor. Appl. Genet. 85: 480-488.

Van Ooijen, J.W., 1992. Accuracy of mapping quantitative trait loci in autogamous species. Theor. Appl. Genet. 84: 803-811.

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

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

Multivariate statistical analysis to assess *Hordeum bulbosum*-mediated haploid production efficiency in barley

T. Adamski, P. Devaux, Z. Kaczmarek & M. Surma, Institute of Plant Genetics, Polish Academy of Sciences, 60-479 Poznań, Poland, and Florimond Desprez Co., POB 41, 59242 Cappelle en Pévèle, France

Key words

barley, haploids, multivariate analysis

Introduction

Barley haploids are often produced by hybridization of *Hordeum vulgare* and *H. bulbosum.* This method requires *in vitro* culture of immature embryos. Haploid production efficiency (HPE) depends on the *H. vulgare* genotype and on environmental conditions. The paper presents results of a multivariate statistical analysis of HPE.

Material and methods

Three spring barley cultivars (Apex, Roland and Vada) and F_1 hybrids derived from crosses of Apex × Roland, Apex × Vada and Vada × Roland were investigated for HPE using the *H. bulbosum* method (Devaux 1986). Apex and Vada have a poor crossability with *H. bulbosum* (Pickering 1993, Devaux *et al.* 1990). In contrast with these two cultivars Roland shows a high crossability with *H. bulbosum.* Interspecific crosses were carried out both in Poland and in France. Seed set (number of seeds per 100 florets), embryo differentiation (number of embryos per 100 seeds), haploid plant development (number of haploid plants per 100 embryos) and haploid plant efficiency (number of haploid plants per 100 florets) were scored and computed. Two-factor multivariate data-analytic methods were used, *i.e.*, cluster analysis, canonical variate analysis and relevant graphical techniques (Caliński & Kaczmarek 1973, Caliński *et al.* 1975, Morrison 1976, Caliński & Corsten 1985).

Results

The general hypothesis about no genotype × location interaction was rejected at the 1% significance level. Therefore estimation and testing of differences between genotypes were done for France and Poland separately. Differences between cultivars and their hybrids were significant with respect to all investigated traits. Using Wilks' likelihood-ratio test (Rao 1973) discriminating power for particular traits was estimated. It was found that haploid production efficiency and seed set were the most discriminating factors for the *H. vulgare* genotypes in both experiments. For these factors three homogeneous groups of genotypes (the same in France and in Poland) were found: 1 - Roland, 2 - Apex and F_1(Apex × Vada), 3 - Vada, F_1(Apex × Roland) and F_1(Vada × Roland). Application of multivariate analysis and related methods allowed a comprehensive representation of information about differences between genotypes regarding factors influencing HPE.

References

Caliński, T. & L.C.A. Corsten, 1985. Clustering means in ANOVA by simultaneous testing. Biometrics 41: 39-48.

Caliński, T. & Z. Kaczmarek, 1973. Metody kompleksowej analizy doświadczenia wielocechowego. Trzecie Colloquium Metodologiczne z Agro-biometrii, PAN, 258-319.

Caliński, T., S. Czajka & Z. Kaczmarek, 1975. Analiza składowych głównych i jej zastosowanie. Roczniki AR w Poznaniu LXXX, ABS 36: 159-185.

Devaux, P., 1986. Yield of haploid production through the *bulbosum* method in a winter barley breeding programme. Cereal Res. Comm. 14: 273-279.

Devaux, P., T. Adamski & M. Surma, 1990. Studies on low crossabilities encountered with the *Hordeum bulbosum* method for haploid production of barley, *Hordeum vulgare* L. Plant Breeding 104: 305-311.

Morrison, D.F., 1976. Multivariate Statistical Methods (2nd ed.). McGraw-Hill Kogakusha, LTD. Tokyo.

Pickering, R.A., 1983. The Location of a gene for incompatibility between *Hordeum vulgare* L. and *H. bulbosum* L. Heredity 51: 455-459.

Rao, C.R., 1973. Linear Statistical Inference and Its Applications (2nd ed.). Wiley, New York.

Variation for grain quality traits in oat (*Avena sativa* L.)

Hermann Buerstmayr, Hermann Joechtl & Peter Ruckenbauer, Institute of Agronomy and Plant Breeding, University of Agriculture Vienna, Vienna, Austria

Key words

Avena sativa, genotype, oats, quality, heritability, variance components

Introduction

The Austrian oat acreage has decreased during the past decades. However, there is an increasing demand for 'high quality oat'. The food processing industry demands white or yellow oat with bright kernel colour, fresh smell, less than 14% moisture, a test weight above 55 kg/100 litres, a groat percentage above 72% and more than 90% of the kernels above 2 mm in diameter. Estimates of the influence of genotype and environment on important oat quality traits are presented.

Materials and methods

In 1993, 36 oat genotypes (19 cultivars and 17 breeding lines) were sown at four locations in Lower-Austria. Three trials were triple lattices and one was a rectangular design with three replicates. The plot size was 10 m^2. After harvest, thousand kernel weight, test weight and sieve fraction above 2 mm were determined. After hulling the groat fraction and thousand groat weight were measured.

Table 1. Overall means, minimum and maximum values, variance components for location, genotype, genotype by location interaction and error as well as heritability estimates

	Mean	Min	Max	σ^2_L	σ^2_G	σ^2_{GL}	σ^2_E	h^2
Thousand kernel weight (g)	35.0	29.6	41.7	3.37	2.93	0.52	3.26	0.88
Test weight (kg/100 litres)	52.3	47.2	56.1	2.86	0.59	0.63	0.79	0.73
% Grains with diameter > 2mm	89.8	65.2	98.5	32.74	10.20	10.46	5.86	0.77
% Groats after hulling	64.8	45.9	70.8	7.79	2.27	15.79	2.87	0.35
Thousand groat weight	26.1	21.2	32.1	7.02	1.50	0.26	1.49	0.89
Grain yield (100 kg/ha)	43.6	27.1	60.1	77.53	1.70	2.51	10.63	0.53

Table 2. Correlation coefficients between quality traits and yield based on overall means (accross four locations) based on 36 oat genotypes

	(1)	(2)	(3)	(4)	(5)	(6)
(1) Thousand kernel weight (g)		0.43^{***}	0.61^{***}	0.32^{***}	0.79^{***}	0.48^{***}
(2) Test weight (kg/100 litres)			0.18^{*}	0.63^{***}	0.49^{***}	$0.02^{n.s.}$
(3) % Grains with diameter > 2 mm				$0.02^{n.s.}$	0.67^{***}	0.65^{***}
(4) % Groats after hulling					0.27^{***}	$-0.05^{n.s.}$
(5) Thousand groat weight						0.65^{**}
(6) Yield						

Results

For an analysis across locations (adjusted) means and a pooled error mean square were computed. Variance components and broad sense heritabilities were calculated (Table 1) as well as phenotypic correlations (Table 2).

Discussion

Heritability was rather high for all quality traits except for groat percentage. Genotype by location interaction for groat percentage was high, but mainly due to one location. An analysis without this location resulted in a heritability of 0.72. Stuthman & Granger (1977) reported heritabilities for groat percentage ranging from 0.34 to 0.72. Location had the most important influence on grain quality as well as on yield.

In contrast to Bunch & Forsberg (1989) correlation coefficients between quality traits and yield were either positive or not significant. Breeding of oat genotypes which combine high yield and grain quality should therefore not be very difficult.

Acknowledgements

We wish to thank Ing. Konrad Schulmeister from Saatzucht Edelhof for the excellent cooperation.

References

Bunch, R.A. & R.A. Forsberg, 1989. Relationships between groat percentage and productivity in an oat headrow series. Crop Science 29: 1409-1411.

Stuthman, D.D. & R.M. Granger, 1977. Selection for caryopsis percentage in oats. Crop Science 17: 411-414.

A computer program for the analysis of the genotype × environment interaction in series of plant breeding experiments

Tadeusz Caliński[1], Stanisław Czajka[1], Zygmunt Kaczmarek[2], Paweł Krajewski[2] & Idzi Siatkowski[1], [1]Dept. of Mathematical and Statistical Methods, Agricultural University, Wojska Polskiego 28, 60-637 Poznań, [2]Institute of Plant Genetics, Polish Academy of Sciences, Strzeszyńska 34, 60-479 Poznań, Poland

Key words
genotype × environment interaction, breeding experiments, computing, statistical analysis

Introduction

The paper contains a description of a computer program for the analysis of a series of experiments with the same genotypes conducted in complete or incomplete block designs in different environments (years or places). The statistical analysis is based on a model described by Caliński *et al.* (1979, 1987), derived from the ANOVA Scheffé-type mixed model. Special attention is given to estimation, hypothesis testing and interpretation of problems concerning genotype × environment interaction.

Variety experiments

For a series of experiments carried out in different environments the following analyses are proposed:

preliminary analysis, consisting of calculating means for varieties in all environments, general means for varieties and environments, and estimates of experimental error,

general analysis, consisting of the analysis of variance and testing of hypotheses concerning environmental effects, main effects of varieties and variety × environment interactions, including regression on the environment,

individual analysis, containing estimation and testing of main and interaction effects and testing of the regression of the interaction on the environment,

analysis of interaction, including calculation of interaction deviations, environmental deviations and decomposition of the F statistic for interaction by means of principal component analysis,

- *analysis of the structure of interaction*, both for varieties and environments, including distances between varieties, between environments, shortest dendrites and the share of interaction attributable to each variety and to each environment.

Plant breeding experiments

In the case of series of plant breeding experiments (*i.e.*, diallel or line × tester), besides the analyses useful for variety testing, the following analyses can be used to investigate general combining abilities (GCA) and specific combining abilities (SCA) of parental entries and heterosis of hybrids (Kaczmarek 1986, Kaczmarek & Krajewski 1991):

- estimation of GCA, SCA and heterosis effects in environments,
- analyses of variance with testing the general hypotheses concerning these effects and their interaction with environments,
- estimation of GCA and SCA for individual parental lines (or pairs of lines) and the analysis of heterosis for hybrids taking into account estimation and testing of the main effects (over environments) and interaction effects,
- analyses of the structure of the GCA × environment, SCA × environment and heterosis × environment interaction by means of the principal components.

The program

In addition to the text output, the program produces graphs illustrating the results of regression and principal component analyses. The course of the analysis can be chosen in an interactive way and saved for later use with the same or similar data. Several parameters of the analyses can be changed. A simple spreadsheet editor is integrated in the program, which allows easy input of data.

References

Caliński T., S. Czajka & Z. Kaczmarek, 1979. On some methods for studying the genotype-environment interaction. Quaderni di Epidemiologia, Suppl. 1: 11-29.

Caliński T., S. Czajka & Z. Kaczmarek, 1987. A model for the analysis of a series of experiments repeated at several places over a period of years. I. Theory, II. Example, Biuletyn Oceny Odmian 13: 34, 35-71.

Kaczmarek Z., 1986. The analysis of a series of experiments in incomplete block designs (in Polish). Roczn. AR Poznań, Rozprawy Naukowe, 155.

Kaczmarek Z. & P. Krajewski, 1991. Statistical analysis of series of experiments with line × tester analysis. Proc. of the Eighth Meeting of EUCARPIA Section Biometrics in Plant Breeding, Brno, 157-165.

Mixture models for mapping QTL without inbred lines

Scott D. Chasalow, Scottish Agricultural Statistics Service, Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA, U.K.

Summary

Models developed for mapping quantitative trait loci (QTL) with inbred lines allow effects of at most two QTL alleles per locus. When inbred lines are unavailable, up to four QTL alleles per locus may be segregating in a mapping population. Use of inbred line models to map QTL in this case may lead to failure to detect QTL, or biased estimation of their effects. We describe a method for mapping QTL without inbred lines that avoids such limitations by using mixture models that allow effects of up to four QTL alleles per locus.

Introduction

When inbred, homozygous parents are unavailable, genetic linkage analysis may proceed using a mapping population originating from a cross between heterozygous parents, but a variety of complications arise. First, more than two alleles at any particular locus may be segregating in a cross. Second, the pair of parental genomes consists of a mosaic of many different genotypic configurations. Finally, linkage phases are generally unknown. Quantitative trait loci (QTL) genotypes are generally unobserved, so complications of linkage analysis with heterozygous parents are even more severe when mapping QTL.

Previous investigators (*e.g.*, Leonards-Schippers *et al.* 1994, Bonierbale *et al.* 1994) have applied models designed for mapping QTL in crosses originating from inbred lines to map QTL without inbred lines. Such "two-component models" allow effects of at most two QTL alleles per locus. When applied to map a QTL for which more than two alleles are segregating, two-component models may fail to detect the QTL or produce biased estimates of its effects if there is interaction among the effects of QTL alleles. In fact, Van Eck *et al.* (1994) recently have reported evidence of three QTL alleles segregating in a diploid potato cross, with interaction among the effects of the QTL alleles. To avoid the limitations inherent in applying two-component models when inbred lines are unavailable, we derived four-component mixture models for QTL mapping that allow effects of up to four QTL alleles per locus. We compare by simulation the performance of a four-component mixture model to that of a two-component mixture model.

Results and discussion

As a simple example, consider the cross, $MQ_1/mQ_2 \times mQ_3/mQ_4$, in which are segregating a single marker with two alleles in a backcross-type configuration, and four distinct alleles of one QTL. A two-component mixture model applicable in this case is $f(y \mid g_k) = P(Q_1 \mid g_k)^*f(y \mid Q_1) + P(Q_2 \mid g_k)^*f(y \mid Q_2)$, $k = 1,2$, where $g_1 = Mm$, $g_2 = mm$, y is the quantitative trait value, and f is a density function. This model includes marginal effects of alleles Q_1 and Q_2, averaging over any effects of Q_3 and Q_4. A four-component mixture model, $f(y \mid g_k) = \sum_i \sum_j P(Q_iQ_j \mid g_k)^*f(y \mid Q_iQ_j)$, $k = 1,2$, $i = 1,2$, $j = 3,4$, allows a different density function, $f(y \mid Q_iQ_j)$, for each of the progeny QTL genotypes, Q_1Q_3, Q_1Q_4, Q_2Q_3, and Q_2Q_4. Such a model allows for interaction between the effect of the QTL alleles from one parent and the effect of the QTL alleles from the other parent. We fit these models by maximum likelihood, adapting the EM algorithm approach for QTL mapping described by Jansen (1992) to estimate simultaneously the recombination fraction and the parameters of the component densities.

The average performance of the two models was compared by fitting both to 100 simulated data sets (Table 1). Each simulation consisted of 200 progeny genotypes randomly generated from the cross, $MQ_1/mQ_2 \times mQ_3/mQ_4$. Each individual was assigned a quantitative trait value randomly sampled from one of four "true" normal distributions, depending on their QTL genotype. The two-component model provided reasonably accurate estimates of the means for the Q_1 and Q_2 components, and a precise estimate of the component SDs. However, this pooled SD estimate is rather biased, since the model assumes the two components have equal variances but the true values are

Table 1. Summary statistics for fit of two- and four-component mixture models to 100 simulated data sets. r is the recombination fraction between the marker locus and the QTL

	Two-component model					Four-component model					
	r	Mean		SD		r	Mean				SD
		Q_1	Q_2	Q_1	Q_2		Q_1Q_3	Q_1Q_4	Q_2Q_3	Q_2Q_4	
True	0.25	25.0	25.0	18.0	11.2	0.25	40.0	10.0	30.0	20.0	10.0
Mean[a]	0.25	25.7	24.6	14.9	14.9	0.17	39.1	11.6	30.3	19.4	10.5
SD[a]	0.02	2.6	2.1	0.7	0.7	0.07	2.1	2.8	3.2	3.4	1.2

[a] - Sample mean and SD of 100 parameter estimates

quite different. The two-component model also provided an accurate estimate of the recombination fraction. As expected, no QTL effects were detected by the two-component model, since there was no true difference between the means of the Q_1 and Q_2 component distributions. The four-component mixture model provided accurate estimates of the four individual component means and the common component SD, but a rather inaccurate estimate of the recombination fraction. With this model, a reasonable test would on average detect the presence of the QTL.

QTL mapping using mixture models requires estimation of the probability of each possible QTL genotype for every individual. Information for this task is provided by the trait value and by linkage to other loci of known genotype. The cross described here, with a single backcross-type marker linked to a QTL, is in some sense a worst-case example. It is well known that single-marker models are inefficient for mapping QTL. In addition, no linkage information is present for distinguishing between QTL alleles Q_3 and Q_4; this distinction relies entirely on information from phenotypic values. We have demonstrated that the four-component mixture model does offer improvements over the previously used two-component models even in this worst-case example, albeit for a QTL with a rather large effect (explains 56% of total variation). Four-component mixture models analogous to that described here may be obtained for other marker configurations and for more than one marker linked to a QTL. Such models, which incorporate relatively more information from linkage, should prove to be even more useful tools for mapping QTL.

References

Bonierbale, M.W., R.L. Plaisted, O. Pineda & S.D. Tanksley, 1994. QTL analysis of trichome-mediated insect resistance in potato. Theor. Appl. Genet. 87: 973-987.

Eck, H.J. van, J.M.E. Jacobs, P. Stam, J. Ton, W.J. Stiekema & E. Jacobsen, 1994. Multiple alleles for tuber shape in diploid potato detected by qualitative and quantitative genetic analysis using RFLPs. Genetics 137: 303-309.

Jansen, R.C., 1992. A general mixture model for mapping quantitative trait loci by using molecular markers. Theor. Appl. Genet. 85: 252-260.

Leonards-Schippers, C., W. Gieffers, R. Schafer-Pregl, E. Ritter, S.J. Knapp, F. Salamini & C. Gebhardt, 1994. Quantitative resistance to *Phytophthora infestans* in potato: a case study for QTL mapping in an allogamous plant species. Genetics 137: 67-77.

Identification of rye grasses using DNA markers

Marc De Loose[1], Kristiaan Van Laecke[1], Ann Depicker[2] & Erik Van Bockstaele[1],
[1]Rijksstation voor Plantenveredeling, Centrum voor Landbouwkundig Onderzoek Gent,
Burg. Van Gansberghelaan 109, B-9820 Merelbeke, [2]Laboratorium voor Genetica,
Universiteit Gent, K.L. Ledeganckstraat 35, B-9000 Gent, Belgium

Introduction

The development of variety specific genetic markers is desirable as an additional tool for variety identification, protection of breeder rights and seed purity determination. Moreover molecular assisted breeding can speed up the time consuming process of developing new varieties.

The genus *Lolium* is an important group of temperate forage grasses, including Italian (*L. multiflorum*) and perennial rye grasses (*L. perenne*). These two species are self-incompatible outbreeding species. The synthetic varieties are often produced by recurrent selection systems. They consist of improved populations that are composed of up to 15 mother plants in the original polycross. Because of the continuing increase in the number of registered varieties, it is becoming more difficult to discriminate all *Lolium* varieties by morphological characteristics. Here we evaluate the RAPD (Random Amplified Polymorphic DNA) markers as an alternative tool to identify rye grasses.

Results

For each variety individual plants were grown in the field. Nuclear DNA was prepared from individual plants. The PCR reactions were essentially performed as described by Williams *et al.* (1990), and the obtained DNA fragments were separated by 2% agarose gelelectrophoresis. The pictures with the DNA profiles were scanned with a HP deskscan II_p. Finally the densitometric RAPD patterns were compared via the GELCOMPAR software (Applied Maths, Kortrijk, Belgium; Vauterin & Vauterin 1992).

Sixty primers were evaluated in a RAPD analysis on an Italian and a perennial rye grass genotype from different origins. All primers yielded defined fragment patterns but only seven of them gave rise to clearly different profiles. These primers were retained for a detailed analysis on 5 Italian and 5 perennial rye grass varieties. For each primer a fingerprint was produced comprising of one to five major bands and a varying number of minor bands. Certain amplified bands appeared to be common to several varieties while others were specific for either perennial or Italian rye grass.

The comparison of RAPD profiles with GELCOMPAR software allowed us to combine the results of different primer driven RAPD reactions in one fingerprint. In this way four RAPD profiles were used to cluster individual plants from different varieties. In the dendrogram, obtained by using the "neighbour joining" clustering method, some varieties are arranged in discrete groups while other varieties are showing an overlap. This is not surprising because existing varieties are often used in new combinations to start a new breeding cycle.

Conclusion

Combining multiple RAPD profiles in the GELCOMPAR software allows to group individual plants of varieties. Therefore, we expect that it will be possible to adapt this strategy for identification and clustering of other plant species that are commercialised as synthetic varieties. We showed that reproducible RAPD polymorphisms allow discrimination between rye grass species. Moreover, most of the individual plants belonging to the same variety are clustered in one group. In the future, species specific RAPD fragments will be cloned and characterised to develop a more specific PCR reaction for routine diagnostic analysis.

References

Vauterin, L. & P. Vauterin, 1992. Computer-aided objective comparison of electrophoresis patterns for grouping and identification of microorganisms. Eur. Microbiol. 1: 37-41.

Williams, J.G.K., A.E. Kubelik, K.J. Levak, J.A. Rafalski & S.C. Tingey, 1990. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res. 18: 6531-6535.

DNA-markers linked to the scab resistance locus introgressed from *Malus floribunda* 821

L. Gianfranceschi, B. Koller, N. Seglias & C. Gessler, Pathology Group, Institute of Plant Sciences, Swiss Federal Institute of Technology, CH-8092 Zürich, Switzerland

Keywords
apple scab, *Venturia inaequalis*, molecular markers, plant breeding

Breeding apple cultivars resistant to scab caused by *Venturia inaequalis* is an alternative way for reducing environmental impact by avoiding fungicide treatments.

The introduction of resistance genes into cultivated varieties could be much improved by molecular marker assisted selection. To find markers linked to the major scab resistance gene, Vf, introgressed from *Malus floribunda* 821, progenies from crosses between resistant and susceptible trees were successfully subjected to Bulked Segregant Analysis (Giovannoni *et al.* 1991, Michelmore *et al.* 1991). Two markers were found and the polymorphic DNA fragments cloned. Transformation of the RAPD markers into more consistent and reproducible markers such as SCARs (Sequence Characterised Amplified Regions) (Paran & Michelmore 1993) and CAPS (Cleaved Amplified Polymorphic Sequences) (Konieczny & Ausubel 1993) were also presented. From the analysis of a segregating population it was possible to calculate that the distance of the markers from the resistance gene was 2.1 and 4.3 cM for $OPM18_{900}$ and $OPU1_{400}$, respectively. The presence or absence of the markers was also tested in some apple cultivars confirming tight linkage with the Vf gene. All tested varieties are in fact carrying the $OPM18_{900}$ marker, and only Coop 13 (Vf-resistant) did not show the presence of $OPU1_{400}$, proving that a recombination event has occurred between the marker and Vf during the breeding process.

The use of these two markers will be very useful in accelerating apple breeding by uncovering rare genetic combinations. Moreover they will also allow simultaneous testing for more than one character.

The present work is part of a wider European apple project. Molecular markers linked to Vf have also been reported by other groups involved in the project. Therefore, a saturated map of the locus will allow accurate segregation analysis of Vf and of the nearby DNA regions. The application of molecular markers in apple breeding will also favour the discovery of functionally different scab resistance genes and, due to the co-dominant character of some of them, it will be possible to reveal homozygotes: that could lead towards the direct transfer of the gene into high quality susceptible cultivars.

References

Giovannoni, J.J., R.A. Wing, M.W. Ganal & S.D. Tanksley, 1991. Isolation of molecular markers from specific chromosomal intervals using DNA pools from existing mapping populations. Nucleic Acid Research 19: 6553-6558.

Konieczny, A. & F.M. Ausubel, 1993. A procedure for mapping *Arabidopsis* mutations using co-dominant ecotype-specific PCR-based markers. The Plant Journal 4: 403-410.

Michelmore, R.W., I. Paran & R.V. Kesseli, 1991. Identification of markers linked to disease-resistance genes by bulked segregant analysis: A rapid method to detect markers in specific genomic regions by using segregating populations. Proc. Natl. Acad. Sci. USA 88: 9828-9832.

Paran I. & R.W. Michelmore, 1993. Development of reliable PCR-based markers linked to downy mildew resistance in lettuce. Theor. Appl. Genet. 85: 985-993.

The use of RAPD markers in the analysis of *Rubus* species

J. Graham, R.J. McNicol & P. Lanham, Soft Fruit Genetics Department, Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA, Scotland

The genus *Rubus* is one of the most diverse in the plant kingdom and has been subdivided into 12 subgenera of which only a few have been domesticated. Molecular analysis may well show that some of the subgenera contain hybrids between species of diverse subgenera and that some of the species have been wrongly assigned to subgenera. Of the domesticated subgenera, the *Idaeobats* contain some 200 species showing considerable differentiation, of which the most important are the European red raspberry (*R. idaeus* subsp. *vulgatus* Arrhen) and the North American red raspberry (*R. idaeus* subsp. *strigosus* Michx) and the black raspberry (*R. occidentalis* L.). The subgenus *Eubatus* is extremely variable and complex. It contains all the blackberries and dewberries and has several sections in South America, a very prominent one in Europe and another in North America. Thousands of taxonomic units have been given specific rank and it is often not possible to assign cultivars to individual species. The *Anoplobatus* contain six species of flowering raspberries which have been used in breeding programmes.

In order to understand the relationships between the various *Rubus* species and also the relationships within species, a study was undertaken using random amplified polymorphic DNA (RAPD) markers. Initially the relationships between the red raspberry cultivars (*R. idaeus*) were examined. Most modern European raspberry cultivars originated from crosses between a few early cultivars such as Newburgh, a *Rubus strigosus* type, Lloyd George and Pynes Royal, both *R. idaeus*, and Preussen, a *R. idaeus* and *R. strigosus* cross. Dale *et al.* (1993) showed that the genetic base of raspberry cultivars released between 1960 and 1988 was becoming narrower with more than 90% having cv. Lloyd George in their pedigrees. The accurate identification of such closely related vegetatively propagated perennial fruit cultivars can be difficult; in the past fingerprinting techniques have proved unsuccessful, using methods such as paper chromatography (Haskell & Garrie 1966) and isoenzyme techniques (Cousineau & Donnelly 1989). Chloroplast DNA probes have also been unable to detect variation between raspberry cultivars (Waugh *et al.* 1990). Recently minisatellite DNA and other oligonucleotide sequences have been used to probe the DNA and produce fingerprints (Nybom *et al.* 1990, Parent & Page 1992). Probing, however, is time consuming and the use of radioisotopes is undesirable. The polymerase chain reaction (PCR) was used for rapid and relatively easy production of fingerprinting patterns in red raspberries. PCR can amplify polymorphic DNA (RAPD) in conjunction with random ten base pair primers (Williams *et al.* 1990). This technique allows differences at the DNA level to be detected by using small amounts of genomic DNA ng-μg as a template for PCR, with a set of random primers, under a specific set of conditions to generate RAPD markers. This study initially investigated whether individual fingerprints could be generated from ten red raspberry cultivars, some closely related, using ten random primers; and to examine how well the relatedness of the cultivars, as shown by similarity indexes produced by analysing the markers generated, matched those produced from their breeding schemes. Ten random primers were used to generate fingerprints with each cultivar being conclusively identified by using three or more random primers (Graham *et al.* 1994).

Due to the success of RAPD markers in fingerprinting closely related cultivars, the same primers were used to examine 14 different *Rubus* species from the three most important subgenera as part of an initial large study on the genus. The markers generated were able to correctly determine the relationships within and between species and group the species into the appropriate sub-genera. One exception was *R. macraeii*, the rare tropical raspberry which has been placed into the *Idaeobats* though this study detected only 26% similarity with the other *Idaeobats* and also with the *Eubats*. A larger study should determine the correct classification of *R. macraeii*. Also the RAPD markers suggested that Black River, thought to be *R. occidentalis*, was probably a *R. occidentalis* × *R. idaeus* cross.

The use of molecular markers may lead to a greater understanding of relationships between species, and more accurate taxonomic classification, as well as to more effective utilisation of genetic diversity by the breeder.

References

Cousineau, J.C. & D.J. Donnelly, 1989. Identification of raspberry cultivars *in vivo* and *in vitro* using isoenzyme analysis. HortScience 24: 490-492.

Dale, A., P.P. Moore, R.J. McNicol, T.M. Sjulin & L.A. Burmistrov, 1993. Genetic diversity of red raspberry varieties throughout the world. J. Amer. Soc. Hort. Sci. 118: 19-129.

Graham, J., R.J. McNicol, K. Greig & M. Van De Ven, 1994. Identification of red raspberry cultivars and an assessment of their relatedness using fingerprints produced by random primers. J. Hort. Sci. 69: 123-130.

Haskell, G. & J.B. Garrie, 1966. Fingerprinting raspberry cultivars by empirical paper chromatography. Journal of Sci of Food Agric 17: 189-192.

Nybom, H., S.H. Rogstad & B.A. Schaal, 1990. Genetic variation detected by use of the M13 'DNA fingerprint' probe in *Malus, Prunus* and *Rubus*. Theor. Appl. Genet. 79: 153-156.

Parent, J.G. & D. Page, 1992. Identification of raspberry cultivars by non-radioactive DNA fingerprinting. HortScience 27: 1108-1110.

Waugh, R., M. Van De Ven, M.S. Phillips & W. Powell, 1990. Chloroplast DNA diversity in the genus *Rubus (Rosaceae)* revealed by southern hybridisation. Plant Systematics and Evolution 172: 65-75.

Williams, J.G.K., A.R. Kubelik, K.J. Livak, J.A. Rafalski & S.V. Tingey, 1990. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nuc. Acid Res. 18: 6531-5.

Further genetic analysis of chickpea (*Cicer arietinum* L.)

E. Hajj Moussa[1]*, T. Millan*[2*]*, J. Gil*[2] *& J.I. Cubero*[2]*,* [1]*Université Saint Joseph, E.S.I.A.M. Zahle, Lebanon,* [2]*E.T.S.I.A.M., Dpto. Genetica, Apdo. 3048, 14080 Cordoba, Spain*

**To whom correspondence should be addressed*

Key words
chickpea, *Cicer arietinum*, gene mapping, isozymes, molecular markers, RAPDs

The importance of genetic maps to accelerate plant breeding procedures has been demonstrated in several crops (Tanksley *et al.* 1984, Stuber & Edwards 1986). In chickpea, a crop which has improved relatively slow, the development of an accurate gene map is indispensable. However, the rapid elaboration of this task faces a major problem: the homogeneity of the genetic background in chickpea, displayed by a low rate of polymorphisms at the molecular level.

In order to obtain a higher rate of polymorphisms, the material under study was obtained from crosses between desi and kabuli types (80 F_6 lines) and from an interspecific cross *C. arietinum* x *C. reticulatum* (40 F_2 plants) differing with regard to qualitative (flower and seed colours, number of pods per peduncle, seed coat thickness) and quantitative characters (days to flowering, seed weight). Twenty-three isozyme systems were assayed (AAP, AAT, ACP, ACO, ADH, DIAP, EP, EST, FK, GAL, GDH, G6PDH, GPI, IDH, LAP, MDH, MPI, NAG, 6PGDH, PGM, PRX, SOD, TPI) using starch gel electrophoresis according to Wendel & Weeden (1990). At the DNA level, 196 primers were surveyed using the RAPD technique (Williams *et al.* 1990). DNA extraction and amplification conditions were applied as reported by Torres *et al.* (1993).

Only 2 isozymes (ADH and GAL) were polymorphic within the F_6 lines and 5 (ACO, ACP, GAL, 6PGDH and GPI) within the F_2 progeny. Thirty primers revealed 38 clear polymorphisms between the F_6 lines with the expected segregation. These primers were also polymorphic in the F_2 progeny.

The small number of polymorphic isozymes in chickpea requires the use of new techniques which may generate a larger number of molecular markers. RAPDs seems to be useful for chickpea gene mapping in view of its advantageous characteristics: high ability in detecting polymorphisms, mendelian segregation, repeatability and easiness of application. The dominant aspect of this kind of markers makes them less efficient in the analysis of F_2 families where it is impossible to distinguish heterozygotes. After analyzing the field data we intend to establish new linkage groups involving morphological markers, isozymes and RAPDs, aimed at extending the gene map of chickpea.

Acknowledgements

This research has been carried out in the Food Legume Unit, CIDA Córdoba. Mr. Hajje Moussa's work has been supported by a grant from the CIHEAM.

References

Stuber, C.W. & M.D. Edwards, 1986. Genotypic selection for improvement of quantitative traits in corn using molecular marker loci. Proc. 41st Annual Corn and Sorghum Research Cnf., Am. Seed Trade Assoc. 41: 40-83.

Tanskley, S.D., C.M. Rick & C. E. Vallejos. 1984. Tight linkage between a nuclear male-sterile locus and an enzyme marker in tomato. Theor. Appl. Genet. 68: 109-113.

Torres, A.M., N.F. Weeden & A. Martín. 1993. Linkage among isozymes, RFLP and RAPD markers in Vicia faba. Theor. Appl. Genet. 85: 937-945.

Wendel, J.F. & N.F. Weeden, 1990. Visualization and interpretation of plant isozymes. In: D.E. Soltis and P.S. Soltis (Eds.). Isozymes in plant biology, pp. 5-45. Dioscorides Press, Portland, Oregon.

Williams J.G.K., A.R. Kubelik, J. Livak, J.A. Rafalski & S.V. Tingey, 1990. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acid Res. 18: 6531-6535.

Application of markers to quantitative analysis in *Lolium*

M.D. Hayward[1], K.G. Hossain[1], N.J. McAdam[1], C.E. Evans[1], J.G. Jones[1], J.W. Forster[2], M. Stammers[2], G.M. Evans[2] & J.K. Will[2], [1]Agricultural and Food Research Council, Institute of Grassland and Environmental Research, Aberystwyth, [2]School of Agricultural Sciences, University of Wales, Aberystwyth, U.K.

The efficiency of selection of quantitative traits in a breeding programme is dependent upon the heritability of the trait concerned, the gene action controlling the trait and the selection intensity imposed. If the character of interest is controlled by loci which are linked to qualitative marker loci, depending upon the degree of recombination between such markers and the quantitative trait locus (QTL), it may well be more efficient to utilize this linkage and select for the marker gene as a means of following the transmission of the trait through generations of selection. Information on the relationship between 'markers' and quantitative traits in the forage grasses may be obtained by several means.

Firstly, by determination of the frequency of specific marker genotypes in populations which have undergone differing degrees of selection for a quantitative trait it may be possible to establish relationships between the presence of the marker and the degree of expression of the trait. A number of populations of *Lolium perenne*, selected for differing quantitative characters, have been analyzed for concomitant changes in the frequency of some specific isozyme alleles. Selection for water soluble carbohydrate for example was accompanied by changes in the frequency of the *b* allele at the PGI/2 locus (Hayward *et al.* 1994). This may indicate a pleiotropic effect of the isozyme locus or linkage between it and a locus affecting the trait.

Secondly, selection for a specific marker and the creation of populations homozygous for differing alleles may result in the establishment of populations which differ for one or more quantitative traits. Populations of *L. perenne* which have been selected for various combinations of isozyme markers have been shown to differ in several important agronomic traits (Hayward *et al.* 1994). The linkage phase for some traits differed between populations thus, this relationship, which again may be due to pleiotropy or linkage, can be utilized for further selection within related populations.

The most effective means of using genetic markers in a breeding programma is when knowledge of the genetic linkage of the quantitative trait to the marker locus, or preferably to loci which flank the QTL on both sides, is well established. The availability of a range of molecular markers, such as isozymes, RFLPs and RAPDs, has enabled detailed genetic maps to be produced. These various approaches to mapping have been applied to two different forms of *Lolium* populations: a family of an F1 hybrid *L. perenne* × *L. multiflorum* backcrossed to a fully homozygous DH *L. perenne* produced by androgenesis and to two sets of double haploid progeny of *L. perenne*. These families have also been assessed in a clonally replicated, fully randomized field experiment for some quantitative traits. To date 150 markers have been analyzed and a map consisting of nine linkage groups created with the aid of Mapmaker. Some difficulties have been encountered in developing the genetic map in the backcross family due to 'map expansion'. This may be accounted for by differences in genome size between the *L. perenne* and *L. multiflorum* parents of the hybrid which may result in structural differences leading to unequal pairing and disturbed segregations. In addition the presence of possible 'pairing genes' (see Evans & Taing Aung 1985) in this parental hybrid may create further problems. The results so far have enabled us to establish nine linkage groups and to identify some loci controlling traits of agronomic importance. The utility of these results for marker assisted selection is being assessed.

References

Evans, G.M. & Taing Aung, 1985. Identification of a diploidizing genotype of *Lolium multiflorum*. Can. J. Genet. Cytol. 27: 506-509.

Hayward, M.D., N.J. McAdam, C. Evans, J.G. Jones, A. Ustin, K.G. Hossein, J.W. Forster, M. Stammers, G.M. Evans & J.K. Will, 1994. Genetic markers and the selection of quantitative traits in forage grasses. Proceedings of the Eucarpia Fodder Crops Section Meeting, Loen, Norway, Aug. 1993, in press.

Detection of molecular markers linked to dominant disease resistance genes: the lod score method revisited with regard to necessary sample sizes

M. Hühn & J. Léon, Institut für Pflanzenbau und Pflanzenzüchtung der Christian-Albrechts-Universität, Olshausenstr. 40, D-24118 Kiel, Germany

Molecular markers have an extremely large potential in the genetic mapping of complex genomes compared to isozyme and morphological markers. Their number is almost unlimited and they are not affected by environmental factors, dominance or epistasis. Some comments on molecular marker-assisted linkage detection for a dominant disease resistance trait of a segregating F_2-population will be given.

The two alleles at the resistance locus are A (= resistant) and a (= susceptible), with A dominant over a. The marker alleles with codominant expression are B_1 and B_2, with recombination value R between marker and resistance locus. Selfing or intercrossing the F_1-genotype AaB_1B_2 of an initial cross of homozygous parents, provides a segregating F_2. Analysis of two-point linkage by the traditional measure of maximum lod score is based on this F_2.

Three subpopulations of the F_2 can be used for linkage analysis: susceptible (= recessive) individuals, resistant (= dominant) individuals and the complete F_2. These are analysed by the traditional approach of maximum lod score: $Z(R) = \log[L(R) / L(0.5)]$. $L(R)$ is the likelihood function depending on the recombination fraction R and log denotes the decimal logarithm.

For the subpopulation of susceptible individuals the expected relative frequencies of the three phenotypically distinct classes are R^2 (for aaB_1B_1), $2R(1 - R)$ (for aaB_1B_2), and $(1 - R)^2$ (for aaB_2B_2). The observed absolute frequencies are denoted by z_1, z_2 and z_3, respectively, with $N = z_1 + z_2 + z_3$. The lod score is:

$$Z(R) = \log\left[R^{2z_1+z_2} (1-R)^{z_2+2z_3} 2^{2N}\right] . \tag{1}$$

It gives the maximum likelihood estimate $\hat{R} = (2z_1 + z_2) / (2N)$. It also leads to the relations $2z_1 + z_2 = 2N\hat{R}$ and $z_2 + 2z_3 = 2N(1 - \hat{R})$. A conventional rule is to conclude that autosomal loci are linked whenever the value of $Z(\hat{R})$ exceeds 3 (Ott 1991). $Z(\hat{R})$

depends on $\hat{R}$, z_1, z_2, z_3 and N. If we replace the frequencies z_1, z_2 and z_3 by expressions depending on N and $\hat{R}$, the condition for a significant linkage, *i.e.*, $Z(\hat{R}) \geq 3$, can be solved for N:

$$F(N,\hat{R}) = 2N\log[2\hat{R}^{\hat{R}}(1-\hat{R})^{(1-\hat{R})}] \geq 3 \ . \tag{2}$$

If we assume linkage with true recombination fraction R, and provided that it is estimated well by $\hat{R}$, then this approach (2) gives a lower bound for the number of individuals N required for detection of significant two-point linkage by the lod score method.

Analogous computations for the subpopulation of resistant individuals (with three phenotypically distinct classes) as well as for the complete F_2 (with six phenotypically distinct classes) require some minor modifications.

Some numerical results are presented in Table 1. With regard to their practical relevance, the numerical sample sizes in Table 1 may be subjected to some criticism since they have been calculated for the special case $R = \hat{R}$.

For sufficiently large sample sizes, $\hat{R}$ is asymptotically normally distributed and unbiased (Ott 1991) with known variance $V(\hat{R})$ which depends on R and N. The previous approach, therefore, can be improved and generalized by the construction of a two-sided central confidence interval about the true recombination fraction R. The resulting necessary sample sizes for this confidence interval can be easily calculated by numerical methods.

For small samples (with unknown variance $V(\hat{R})$), however, the limits $\hat{R} \pm \surd V(\hat{R})$, with $\surd V(\hat{R}) = vR$, where v denotes the coefficient of variation for $\hat{R}$, may provide some rough numerical results on necessary sample sizes for linkage detection. Some results are presented in Table 2. The exact distribution of $\hat{R}$ for small samples and, therefore, exact confidence intervals and necessary sample sizes are just derived by simulation.

Table 1. Lower bounds of necessary sample sizes required for linkage detection

Subpopulation	Recombination fraction								
	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45
Susceptible	7	10	13	18	27	42	76	172	690
Resistant	39	57	84	128	201	338	638	1500	6165
Susc. + Res.	19	26	36	51	76	123	223	511	2064

Table 2. Necessary sample sizes for different values of the recombination value, *R*, and the coefficient of variation of $\hat{R}$, *v*

v	*R*			
	0.05	0.15	0.25	0.35
susceptible individuals:				
0.05	7 - 8	13 - 14	24 - 30	61 - 98
0.10	7 - 8	12 - 15	22 - 33	50 - 130
0.20	7 - 8	11 - 16	18 - 43	35 - 268
0.30	7 - 8	10 - 18	16 - 56	26 - 883
resistant individuals:				
0.05	38 - 40	80 - 90	178 - 227	503 - 826
0.10	37 - 40	75 - 95	159 - 258	404 - 1112
0.20	36 - 42	67 - 108	127 - 338	272 - 2369
0.30	34 - 44	59 - 122	103 - 455	192 - 8182
susceptible and resistant individuals:				
0.05	18 - 19	34 - 37	69 - 85	178 - 288
0.10	18 - 19	32 - 39	62 - 96	145 - 386
0.20	17 - 20	29 - 44	51 - 123	100 - 811
0.30	17 - 20	26 - 49	42 - 162	53 - 2609

Reference

Ott, J., 1991. Analysis of human genetic linkage. Revised edition. The John Hopkins University Press, Baltimore and London.

A system for statistical analysis of genetic and breeding experiments

R. Kala, H. Chudzik, A. Dobek & H. Kiełczewska, Department of Mathematical and Statistical Methods, Agricultural University of Poznań, Poland

Key words

computer program, block design, diallel cross

The system is devoted to the analysis of data obtained from genetical and/or breeding experiments, in which the set of parental lines and a chosen crossing system are essential elements. The system includes various univariate statistical analyses for experiments with genotypes obtained in one of the four Griffing's types of diallel crossings (Griffing 1956). It takes into account both the completely randomized designs and any block designs. The range of allowed block designs covers the classic completely randomized blocks as well as binary incomplete blocks or any overcomplete block designs. The only restriction is the connectedness of the design, which ensures estimability of any treatment comparison and, in result, any genetic effect.

The system recognises the type of diallel cross after indicating the classifications connected with parental lines. For each of the diallels the genetic analysis provides:

a. the tables of means for parental lines and hybrids,

b. the extended analysis of variance, *i.e.*, the analysis in which the variance between hybrids is divided into variances corresponding to general combining ability, specific combining ability, and, in the case of diallel type I and III, reciprocal effects,

c. the estimates of various breeding parameters, *i.e.*, the general combining abilities, the specific combining abilities, and reciprocal effects, if possible.

For each of the estimates the significance level of the corresponding test statistic is given. The tests for the differences between pairs of breeding values are also provided. In the case of diallel types I and II the system evaluates the effects of heterosis of the progeny in relation to the parental lines, of the hybrids in relation to the midparent, and of the hybrids in relation to the better parent.

For diallels of types I and II the system provides also full characterisation of parental lines with respect to the dominant or recessive gene action. First, the usual assumption of diploid segregation, of independent action of non-allelic genes, of no multiple allelism, of parents homozygosity, and of independent distribution of genes, are tested. If these assumptions are satisfied the full analysis of variance concerning the gene action is calculated. It enables to test the significance of additive gene action as well as dominance. In the presence of dominance the hypothesis of one direction dominance (for diallels of types I and II) and of symmetry in gene distribution (only for diallel type I) can be verified. Finally, Mather's parameters (Mather & Jinks 1982) are estimated and then the basic genetical characteristics, i.e. the mean level of dominance over all loci, the number of gene groups exhibiting dominance, the ratio of the total numbers of dominant to recessive genes in all parents, the coefficients of heritability in broad and narrow sense, are calculated (Allard 1960, Falconer 1970, Dobek *et al.* 1989).

The system enables also the analysis of simple experiments, in which any set of treatments, not necessarily parental lines or offsprings, are compared. In this case the system provides the table of means for treatments adjusted with respect to blocks, the analysis of variance table and the analysis of contrasts. It is also possible to define specific contrasts and to receive their estimate together with the corresponding test statistics.

The system is equipped with an editor, a help subsystem, and provides full menu-mouse communication in the Turbo Pascal 7.0 environment.

References

Allard, R.W., 1960. Principles of Plant Breeding. Wiley, New York.

Dobek, A., Z. Kaczmarek, H. Kiełczewska & T. Łuczkiewicz, 1989. Genetic analysis of a half diallel. Listy Biom. 26: 21-28.

Falconer, D.S., 1970. Introduction to Quantitative Genetics. Longman, New York.

Griffing, B., 1956. Concept of general and specific combining ability in relation to diallel crossing systems. Australian J. Biol. Sci. 9: 463-493.

Mather, K. & J.L. Jinks, 1982. Biometrical Genetics (2nd ed.). Chapman and Hall, London.

Use of molecular markers to locate quantitative trait loci in barley

B. Kjær & J. Jensen, Risø National Laboratory, Roskilde, Denmark

Most characters of agricultural crops show continuous variation. In spite of the importance of these characters, the knowledge of their genetic bases is poor. The inheritance is complex, usually assumed to involve numerous genetic factors that frequently interact with environmental effects. Generally, the genetic factors have not been identified individually, and little is known about the quantitative trait loci (QTLs) which lead to continuous variation. However, the use of mapped genetic markers provides a powerful approach for studying quantitative traits and for locating individual genetic factors associated with the trait. Genetic linkage maps make it possible to evaluate the entire genome for QTLs (*cf.* Tanksley 1993). By using molecular markers in barley QTLs have been found for agronomic traits, malting quality (Hayes *et al.* 1993a), quantitative powdery mildew resistance (Heun 1992), winter hardiness (Hayes *et al.* 1993b) and milling energy (Chalmers *et al.* 1993).

In this study a genetic analysis was carried out on 79 DH lines produced by the *Bulbosum* method from F_1-plants of a cross between the 2-rowed spring barley 'Tystofte Prentice' and the 6-rowed winter barley 'Vogelsanger Gold'. The presence of polymorphisms in the two parent varieties and in the DH lines was studied with regard to 85 markers including 4 morphological markers, 8 isozymes, 70 RFLPs and 3 RAPDs. The agronomic traits, including heading date, stem length, grain yield, thousand grain weight, number of kernels per spike and number of spikes per m^2, were recorded in 1989 and 1991 in a field trial with 'T. Prentice' and 79 DH lines. Only DH lines with no vernalization requirement were used in the trial. The design of the experiment was an incomplete block design with 4 replicates. The quantitative traits were analyzed by the interval mapping approach with Mapmaker/QTL (Lander & Botstein 1989) and multiple regression.

Between 1 and 3 QTLs were found for each of the traits. A large part of the QTLs was found in 1989 as well as in 1991, with similar positions on the genome. Most QTLs were clustered in two areas on chromosome 2. QTLs affecting different traits fell near one another more frequently than would be expected by chance. This suggest that the observed correlations between traits partly may be due to pleiotropic effects of single

QTLs. Together, the QTLs explained 70-88% of the variation between DH-lines for heading date, thousand grain weight and number of kernels per spike, whereas for grain yield only 17-34% of the variation was explained. For all traits, except for grain yield in 1989 and for the number of spikes in 1991, a QTL was found near locus *v* (2-row/6-row) on chromosome 2. A pleiotropic effect of locus *v* could not be rejected. QTLs near locus *Xris39b* on chromosome 2 were in the same chromosomal area as the QTLs identified by Hayes *et al.* (1993a) in a cross of two 6-rowed barleys for heading date, stem length and grain yield. The QTLs for thousand grain weight on chromosome 4 and at locus *v* on chromosome 2 showed significant interaction. The QTL on chromosome 4 for thousand grain weight was only expressed in the 2-rowed lines.

References

Chalmers, K.J., U.M. Barua, C.A. Hackett, W.T.B. Thomas, R. Waugh & W. Powell, 1993. Identification of RAPD markers linked to genetic factors controlling the milling energy requirement of barley. Theor. Appl. Genet. 87: 314-320.

Hayes, P., B.H. Liu, S.J. Knapp, F.Q. Chen, B. Jones, T.K. Blake, J.D. Franckowiak, D. Rasmusson, M. Sorrels, S.E. Ullrich, D. Wesenberg & A. Kleinhofs, 1993a. Quantitative trait locus effect and environmental interaction in a sample of North American barley germ plasm. Theor. Appl. Genet 87: 392-401.

Hayes, P.M., T. Blake, T.H.H. Chen, S. Tragoonrung, F. Chen, A. Pan & B. Liu, 1993b. Quantitative trait loci on barley (*Hordeum vulgare* L.) chromosome 7 associated with components of winterhardiness. Genome 36: 66-71.

Heun, M., 1992. Mapping quantitative powdery mildew resistance of barley using a restriction fragment length polymorphism map. Genome 35: 1019-1025.

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

Tanksley, S.D., 1993. Mapping polygenes. Annual Review of Genetics 27: 205-233.

Biochemical markers for the identification of hop (*Humulus lupulus* L.) germplasm

D. Kralj[1], *Dj. Vasilj*[2] *& M. Kač*[3], [1]*Institute for Hop Research and Brewing, 63310 Žalec, Slovenija,* [2]*Faculty of Agriculture, 41000 Zagreb, Croatia,* [3]*Biotechnical Faculty, 61000 Ljubljana, Slovenija*

Hop (*Humulus lupulus* L.) essential oil consists of numerous components which can best be detected by gas chromatography. Its composition varies very little from year to year but may differ considerably in different accessions. Therefore, hop essential oil can be used to describe the genetic variation of hops, *i.e.*, to identify its germplasm. More than 300 components can be detected in a hop essential oil (Moir 1992), the quantities of which vary from trace amounts to up to 65% (Kralj 1991). The presentation of data has been done in many different ways. Because of high positive correlations between some components there is no need to consider all peaks when evaluating a chromatogram. Usually only a limited number of components are taken into account and only some characteristically different cultivars are considered. As we wanted to observe all possible genetic variation, not only modern, but also primitive accessions were included in our study.

Essential oils of 95 different accessions originating, from different hop growing districts worldwide, were studied in five successive years; 187 components were quantified. Data were analyzed by multivariate methods, whereby the results obtained by rotation-factor analyses were the most meaningful. To describe the variation of the essential oils 31 parameters were chosen (30 components and the α-humulene / β-cariophyllene ratio). The results were classified using a M_{min}-M_{max} matrix. This enabled us to take care of the year-to-year variations in the essential oil composition as well as minor variations within the same type of oil. The essential oils of 95 accessions were thus grouped into 14 groups which reveal the basic genetic variation of hops.

As long as various ingredients are given by means of relative percentages the characteristics (biochemical markers) linked to more than one component could go unnoticed. The relative percentages can differ by two or three orders of magnitude, so they cannot be compared directly. We assumed that some genetic differences and similarities, characteristic for various ecotypes, could be linked to various ratios between some crucial components of the essential oils. In order to get these components "on the same scale" the fraction of each essential oil component was not given as a relative percentage of each substance but by an index (X_N) denoting the relative percentage of the substance in question compared to its maximum content (the maximum content for each component means an index of 100):

$$X_N = \frac{\text{relative \% of component } N \text{ in a given sample}}{\text{maximum relative \% of component } N \text{ in all the samples studied}} \times 100 \ .$$

In order to make the comparison easier, European traditional aromatic hops (English Fuggles and Goldings, Czech hops and Bavarian hops) were further chosen as a model group, *i.e.*, as reference oils. The thirty ingredients (31 parameters) were divided into obligatory ingredients (which are characteristic for all hop oils in the reference group) and facultative ingredients (which occur only in some oils in the group). The ingredients may be facultative in the sense that they are never present in the essential oils of some accessions or in the sense that they do not occur every year. The ingredients of both groups were then subdivided according to descendent indexes. Four groups of components were formed: obligatory components with an index above 50, obligatory components with an index below 50, facultative components with an index above 50 and facultative components with an index below 50.

Totals of indexes for each oil, their subtotals for each subgroup, the ratios of the subtotals for the first and the second subgroups and those for the first and the fourth subgroups are characteristically different for various genotypes. Additionally, one can consider the number of components with very high or very low indexes. This method also revealed some very interesting ratios for various crucial components which are also significantly different for various genotypes and some that can be further linked even to aroma score (*e.g.*, the ratio α-humulene : δ-cadinene : geranyl acetate is very useful in both cases).

The method gives a thorough insight into the genetic variation of hops as reflected in the composition of the hop essential oils. Results can also be used to reveal the duplication of names for some accessions. By this method not only the type of the essential oil but also each single accession can be identified. This can be of importance for identification of old, well established accessions as well as for identification of new cultivars with modified germplasm.

References

Kralj, D., 1988. Variability of Essential Oils of Hops, *Humulus lupulus* L. Journal of the Institute of Brewing 97: 197.

Moir, M., 1992. The 1990 Laurence Bishop Silver Medal lecture. The desideratum for flavour control. Journal of the Institute of Brewing 98: 215.

Genetic variation in blackcurrant (*Ribes nigrum* L.) detected by molecular markers

P.G. Lanham[1], R.M. Brennan[1], C. Hackett[2] & R.J. McNicol[1], [1]Soft Fruit Genetics Department, [2]Scottish Agricultural Statistics Service, Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA, UK

The genus *Ribes* consists of approximately 150 species found mainly in northern temperate regions of Europe and North America (Brennan 1990). The blackcurrant, (*Ribes nigrum* L.) is the most commercially important species in Europe, with 3200 ha grown in the UK alone (Brennan *et al.* 1993). There is considerable scope for genetic improvement of the blackcurrant with respect to low-temperature hardiness, resistance to pests such as gall mite (*Cecidophyopsis ribis* Westw.) and properties relating to fruit quality (*e.g.*, anthocyanin content, yield, machine harvesting ability, etc.). It is for these reasons we have developed and are now using molecular markers to characterise the genetic variation present in *R. nigrum* germplasm.

Twenty-one genotypes, representing a broad spectrum of available germplasm (including European, Scandinavian and Russian subgroups, cultivars developed directly from wild accessions and seedlings derived from interspecific hybridisations), were screened for random amplified polymorphic DNAs (RAPDs, Williams *et al.* 1990), resulting in the identification of 54 markers (Lanham *et al.* 1994). Each of the 21 genotypes could be distinguished using these markers. However, these 54 markers represented 26% only of the total number of amplification products which were scored, indicating a relatively narrow genetic base has hitherto been used for blackcurrant improvement.

The detection of variation by other means was considered desirable. Initial studies on a limited number of *R. nigrum* cultivars using (GATA)$_4$ as an 'RFLP-type' probe confirmed that such variation could be detected using microsatellites. A broader study encompassing 21 genotypes used in the RAPD experiment and additional microsatellite sequences will reveal the full extent of this variation. However, the usefulness of microsatellite markers may be limited either by the amounts of DNA required for RFLP-type experiments or the amount of time required for cloning and sequencing microsatellite containing regions to design primers for their detection using the polymerase chain reaction.

Wild *Ribes* species constitute an important genetic resource for blackcurrant improvement. Variation among wild species was detected using RAPD markers and was found to be more extensive than that found in *R. nigrum* genotypes, for example the most informative primer used with *R. nigrum* was OPA-06 (GGTCCCTGAC, Operon Technologies) which generated six RAPDs, whereas OPA-20 (GTTGCGATCC, Operon Technologies) generated 23 RAPDs among 22 wild species.

The results of these experiments demonstrate that molecular markers detect genetic variation in *Ribes* germplasm at the level of DNA sequence and indicate the potential of this technology in fingerprinting studies, phylogenetic analysis, genetic mapping experiments and the identification of marker tagged traits of agronomic importance.

References

Brennan, R.M., 1990. Currants and gooseberries (*Ribes*). *In:* Genetic resources of temperate fruit and nut crops. Moore, J.N. & J.R. Ballington (Eds.), pp. 457-488, ISHS Publications, the Netherlands.

Brennan, R.M., P.G. Lanham & R.J. McNicol, 1993. *Ribes* breeding and research in the UK. Acta Hort. 352: 267-276.

Lanham, P.G., R.M. Brennan, C. Hackett & R.J. McNicol, 1994. RAPD fingerprinting of blackcurrant (*Ribes nigrum* L.) cultivars. Theor. Appl. Gen.: in press.

Williams, J.G.K., A.R. Kubelik, K.J. Livak, J.A. Rifalski & S.V. Tingey, 1990. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nuc. Ac. Res. 18: 6531-6535.

Inheritance of morphine content in a diallel cross of poppies (*Papaver somniferum* L.)

K. Lőkös-Tóth & E.L. Heszky, Department of Genetics and Plant Breeding, Gödöllő University of Agricultural Sciences, H-2103 Gödöllő, HUNGARY

The aim of this paper is to investigate the inheritance of morphine content of poppies (*Papaver somniferum* L.) by means of a diallel cross. The experiment was carried out at the Experiment Station of Gödöllő University of Agricultural Science during 1992-1993. The F_1 generation from a diallel cross of five diverse *Papaver somniferum* genotypes and cultivars was evaluated for morphine content and some morphological characters of the poppy capsule (length of capsule, width of capsule, number of compartments). General and specific combining abilities were studied by the methodology of Griffing (1956). Genetic components were calculated according to Hayman (1954). Graphical analysis of variance and covariance was done according to Jinks & Hayman (1954).

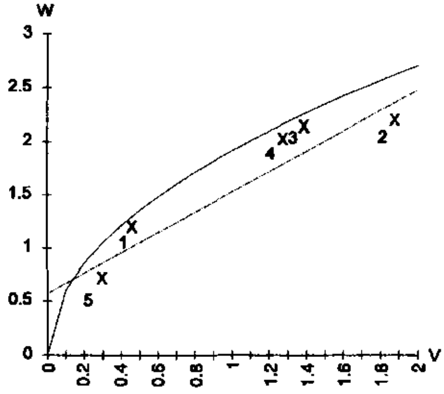
Despite the fact that the parental lines show quite big differences in morphine content (from 3.81% to 8.88%) the F_1 generations did not show much higher values (Table 1). Moreover, the variance of the F_1 generations decreased compared to the variance of the five parents. This is probably due to a more uniform distribution of genes in the F_1s and a lack of heterosis.

Analysis of combining abilities showed that the variance of the general combining ability (GCA) was significant at the 0.05 level but that the variance of specific combining ability (SCA) was not significant. Although reciprocal effects were also significant, its contribution was small compared to that of GCA. This indicated that the additive genetic effect is most important in the genetic control of morphine content. The correlation between the morphine content of parents and GCA effect values of parents is

Table 1. Morphine contents of 5 parents and their combinations in the F_1 generation

Parent	1.	2.	3.	4.	5.	Mean
1. Kompolti mák	8.88	7.45	8.61	9.35	8.72	8.60
2. T - 2	7.15	3.81	5.81	5.77	5.69	5.65
3. B - 1	7.76	4.42	7.60	7.48	6.13	6.68
4. Kék Duna	8.31	5.96	6.30	7.73	7.22	7.10
5. K. rezisztens	6.71	7.73	8.96	5.92	6.70	7.20
Mean	7.76	5.87	7.46	7.25	6.89	

Figure 1. Graphical analysis of the inheritance of morphine content in the F_1 generation



quite high (r = 0.96). This provides a solid basis for choosing parents in a breeding program.

The maximum value obtained for heritability in narrow-sense was 0.76, as calculated from components of variance.

The W-V graph supported the fact that the additive genetic effect plays a great part in the inheritance of morphine content (Figure 1). The points of the parents, except for parent 5, are located equally along the regression line according to their order of dominance. The parents containing the most dominant genes are nearest to the origin. The correlation between the parental order of dominance and parental value of morphine content was −0.47. This means that parents containing the most favourable genes contain the most dominant genes as well.

The variation due to additive effects was 3.42 and the variance due to dominance effects was 0.95. The mean degree of dominance was 0.53. The proportion of genes with positive and negative effect in the parents was 0.25. The proportion of dominant and recessive genes in the parents was 1.47.

The morphine content showed correlation only with the capsule width (r = −0.49) in the characteristics of capsule investigated.

References

Griffing, B., 1956. Concept of general and specific combining ability in relation to diallel crossing systems. Australian Journal of Biological Sciences 9: 463-493.

Hayman, B.I., 1954. The theory and analysis of diallel crosses. Genetics 39: 789-809.

Jinks, J.L., 1954. The analysis of continuous variation in a diallel cross of *Nicotiana rustica* varieties. Genetics 39: 767-788.

Mapping molecular markers showing segregation distortions

M. Lorieux[1], F. Luro[2], B. Goffinet[3] & X. Perrier[4], [1]CIRAD-BIOTROP and [4]CIRAD-FLHOR, B.P. 5035, F-34032 Montpellier Cedex 1, [2]Laboratoire de Biologie Cellulaire et Moléculaire, INRA Bordeaux, B.P. 81, F-33883 Villenave d'Ornon Cedex, [3]Station de Biométrie et d'Intelligence Artificielle, INRA, B.P. 27, F-31326 Castanet-Tolosan Cedex

Key words

segregation distortions, molecular markers, mapping, maximum likelihood estimates, *Citrus*

Introduction

We present a mapping method which takes into account deviations from single-locus segregation ratios, so called segregation distortions. The method is briefly described for backcross and F_2 populations, with an application to the genetic mapping of *Citrus.*

Methods

Consider two markers, A and B, which show segregation distortions. Suppose that these distortions are induced by viability differences between gametes or zygotes due to one or more selected alleles. Then, it is possible to write the expected frequencies of the phenotypic classes as a function of *r*, the recombination frequency between loci A and B and *u* and *v*, the parameters which represent the viability of genotype 1 *vs* genotype 2 at locus A and B, respectively (backcross case). Maximum likelihood estimates (MLEs) can be derived from these expected frequencies. Similarly, it is possible to write likelihoods for the case of several markers.

Backcross populations

Testing for linkage

We have shown that in the case of segregation distortions, classical statistics (χ^2, LOD score) used for testing linkage may lead to the grouping of markers that are not linked

(Lorieux *et al.* 1994b). As an alternative, we suggest using a LOD score which takes the distortions into account.

Estimating linkage

Similarly, we have shown that classical estimates of recombination frequencies may be strongly biased in the case of segregation distortions. Bailey's estimate (Bailey 1949) was found to be consistent and efficient under more general assumptions than those defined by its author, even if the alleles responsible for selection are not located on the marker loci. This estimate should therefore be used instead of the classical estimate.

Ordering markers

We have also shown that the order of the markers in a linkage group may be affected by segregation distortions, when a classical multi-point analysis is used. We propose an alternative method for ordering such markers (see section "Application").

F_2 populations

For F_2 populations, gametic and zygotic selection may affect the analysis of linkage in different ways. Therefore, specific likelihood equations have to be developed for each case, including dominant and codominant markers. The asymptotic bias of the classical estimates were derived for each case, in order to compare them with the standard deviations of the suggested estimates. For each situation MLEs were derived. These should be used instead of classical estimates. We have shown that dominant markers provide little information in case of segregation distortions, and therefore should be used with circumspection (Lorieux *et al.* 1994a). The precision of estimates of recombination frequencies is less affected by selection for codominant markers than for dominant markers.

Application

The models developed here were applied to build a map using a cross between a *Citrus grandis* and a hybrid of a *Citrus reticulata* and a *Poncirus*. We found significant differences between the map obtained by classical methods (MapMaker, Lander *et al.* 1987) and by the method suggested here (Luro *et al.* 1994). A detailed analysis of the results allowed us to conclude that the correction was correct.

Discussion

Segregation distortions should be taken into account in each map construction. The methods presented here and described by Lorieux *et al.* (1994a, 1994b) should be useful for this aim. Other possible sources of segregation distortions, *e.g.*, structural rearrangements such as translocations were not considered, because the answer is probably not a statistical one.

References

Bailey, N.T.J., 1949. The estimation of linkage with differential viability, II and III. Heredity 3: 220-228.

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

Lorieux M., B. Goffinet, X. Perrier, D. González de León & C. Lanaud, 1994a. Maximum likelihood models for mapping genetic markers showing segregation distortions. 1. Backcross populations. Theor. Appl. Gen.: in press.

Lorieux M., X. Perrier, B. Goffinet, C. Lanaud & D. González de León, 1994b. Maximum likelihood models for mapping genetic markers showing segregation distortions. 2. F_2 populations. Theor. Appl. Gen.: in press.

Luro F., M. Lorieux, F. Laigret, J.M. Bové & P. Ollitrault, 1994. Cartographie du génome des agrumes à l'aide des marqueurs moléculaires et distorsions de ségrégation. To be published in: "Techniques et utilisations des marqueurs moléculaires", 29-31 march 1994, Montpellier, France.

Genetic analysis of quantitative traits of oil sunflower and its application to breeding

T. Łuczkiewicz, Department of Genetics and Plant Breeding, Agricultural University of Poznań, Poland

Key words

diallel cross, earliness, oil content

Genetic and breeding studies in oil sunflower were carried out from 1967 to 1989. In the first phase of the investigations new genetic variation was induced using Röntgen radiation in two oil sunflower varieties, Czernianka 66 and Karlik 68. The achenes of these varieties were irradiated in two generations with doses varying from 5 to 25 kR. Forty-eight lines out of 900 inbred lines were chosen in the early generation and eight in the late generation.

By using canonical variate analysis the eight lines, chosen from the 48 lines examined in the early generation, showed the largest variation with regard to sixteen traits. Diallel crosses were performed between the eight lines according to Griffing's model I (Griffing 1956). Experiments were done in two years using the F_2 generations obtained from the original crosses. The experiments were carried out as 8 × 8 lattice squares with four replicates. Measurement and observations of the following plant traits were conducted: seedling height (plant height in the stage of the first pair of leaves), final plant height, number of leaves, flowering period, days from emergence to the beginning of flowering, head diameter, number of seeds per plant, achenes weight per plant, 1000 achenes weight and oil yield per plant.

Genetic determination of quantitative traits could be estimated for nine out of eleven examined traits. For 1000 achenes weight and oil yield per plant the additive-dominant model appeared to be non-adequate in both experiments. For five traits a genetic analysis could only be performed in one year because of significant maternal effects (for days from emergence to beginning of flowering and seed weight per plant) or non-adequacy of the model (for head diameter and 1000 achenes weight in the first year, and plant height in the second year). In the case of revealing the adequacy of additive-dominant model uni-directional effects of dominant genes were statistically significant (with the exception of flowering period for which ambi-directional dominance occurred).

For oil content in achenes the additive variance was in the first year four times larger than variation connected with dominance effects and six times larger in the second year. The above results are similar with those obtained, *e.g.*, by Kadkol *et al.* (1984). For days from emergence to the beginning of flowering the variation for additive effects was also six times bigger than the variation for dominance effects. It confirms results of Manjunath & Goud (1983).

For the majority of the investigated traits of the F_2 generations asymmetric gene distributions with regard to parental forms were obtained.

The ratio of dominant genes to recessive genes was changing depending on the year of the experiment, genotype and the trait. Relatively best accordance of the results in both years of experiments was obtained for oil content in achenes. Environmental conditions had the greatest effect on the estimate of the ratio of dominant to recessive genes for flowering period.

The analysis of dominant to recessive genes ratio allows the statement that in the eight examined inbred lines of sunflower dominant genes outnumber recessive genes 2-3 times. Taking into account genetic determination of examined traits and their variation, plants revealing transgression for oil content in achenes and days from emergence to the beginning of plant flowering, were selected in the F_2 generation of the diallel crosses. Individual selection carried out in nine generations (F_2-F_{11}) yielded genotypes with a short period from emergence to the beginning of flowering (fourteen days earlier than the control) and genotypes with a high oil content in the achenes (6-7% more than the control).

References

Griffing, B., 1956. Concept of general and specific combining ability in relation to diallel crossing system. Austr. J. Biol. Sci. 9: 463-493.

Kadkol, G.P., J. Anand & R.P. Sharma, 1984. Combining ability and heterosis in sunflower. Indian J. Genet. and Plant Breeding 44: 447 - 451.

Manjunath, A. & J.V. Goud, 1983. Genetics of quantitative characters in sunflower (*Helianthus annuus* L.). Genet. iber. 35: 13-23.

Classification of locations in sugar beet trials

P. Müller[1], W.E. Weber[2], G. Steinrücken[3] & G. Diener[4], [1]Institut für Angewandte Genetik, Universität Hannover, [2]Institut für Pflanzenzüchtung und Saatgutwirtschaft, Martin-Luther-Universität Halle-Wittenberg, [3]A. Dieckmann-Heimburg, Nienstädt, [4]KWS, Einbeck

Performance trials in sugar beet (*Beta vulgaris* L.) are conducted by breeders at many locations distributed over the whole area, for which varieties are developed. Such experiments are very expensive. Breeders are interested in reducing the number of locations with minimum loss of information. However, large interactions of genotypes with locations and/or years raise problems.

In this study ways are investigated to find regions of similar response, so that interactions within regions are small. For that purpose performance trials of two important breeding companies in Europe have been analysed biometrically. The data base is very large and includes many series of experiments over two years. One series consists of a set of genotypes tested at several locations. For each set of locations more than one series exists so that results can be checked by cross-validation.

Like in other cases, cross-validation was not overwhelming. Therefore classification methods were made to test if the classification was purely random (see Sneath & Sokal 1973, Bock 1974). For that purpose two parameters were developed. The first parameter describes the stability of a classification, based on distance measures between pairs of locations. Ideal would be the same distance in each series. The second parameter describes the structure by measuring the increase of heterogeneity with continued fusing of clusters. The goal would be to find two or three homogeneous clusters with much heterogeneity between clusters. Again several series can be compared.

For both parameters statistical tests have been developed based on simulation studies. If no real classification exists, the distribution of the parameters is developed by simulation. Several of the obtained classifications were not random. The best distance measure was the Euclidean distance, the best cluster method was complete linkage. The results varied for different traits. Sodium content showed the most clear results. Hierarchical classification methods do not allow incomplete classification, which may be more appropriate for sugar beet. Incomplete classification will be studied next.

Another method to look at locations is to use the regression approach (Weber & Vanselow 1985, Weber & Wricke 1990). With this approach good locations show a steep slope together with no variation around the regression line. Such locations are preferred for selection, since selection of genotypes is safer. It is important to know whether the slope is characteristic for a location over series of experiments. The analysis of the sugar beet data revealed that large differences between slopes exist. Regression coefficients were positively correlated with the residual variance, if unstandardized observations were used.

References

Bock, H.H., 1974. Automatische Klassifikation. Vanderhoeck & Ruprecht, Göttingen

Sneath, P.H.A. & R.R. Sokal, 1973. Principles of Numerical Taxonomy. Freeman, San Francisco.

Weber, W.E. & M. Vanselow, 1985. Die Eignung von Prüforten zur Selektion von Sorten auf Ertrag, ermittelt aus amtlichen Prüfungen bei Winterweizen und Mais. Z. Pflanzenzüchtg. 94: 64-73.

Weber, W.E. & G. Wricke, 1990. Genotype × environment interaction and its implication in plant breeding. In: M.S. Kang (Ed.), Genotype - Environment Interaction and Plant Breeding, pp. 1-19. Louisiana State Univ, Baton Rouge, U.S.A..

Efficiency of check plots to control the soil heterogeneity in field trials

Josef Pešek, Research Institute of Animal Nutrition Pohořelice, Department of Soil Management, 664 62 Hrušovany near Brno, Czech Republic

The efficiency of check plots in field trials

Theoretical considerations based on Smith's (1938) coefficient of soil heterogeneity suggest that the efficiency of systematically arranged control plots in field trials depends upon the type and degree of soil heterogeneity at an experimental site, on the span of the check plots, and on the way the adjustment of yields of test plots is carried out. Data from uniformity trials with spring wheat (*Triticum aestivum* L.) and with peas (*Pisum sativum* L.) support the theoretical conclusions (Pešek 1973). The reinforcement of field trials by widely spanned control plots may reduce the error variance markedly, if the fertility index is used as a concomitant variable for adjusting yields of test plots by analysis of covariance. Mapping experimental areas according to Smith's coefficient of soil heterogeneity could be useful for the choice of the optimal experimental strategy for a given field.

The efficiency of control in BIB designs (Pešek 1974)

The variances of treatment differences for balanced incomplete block designs, having an extra control added to each block, are presented. The efficiency of such a design was compared with the corresponding conventional designs as given by Cochran & Cox (1957). The addition of and extra control to balanced incomplete block designs can be recommended if the comparison of each treatment with the control is the goal of the field trial. However, the loss of precision in treatment comparisons is appreciable in experiments with a small number of replications. The precision of treatment versus control comparisons increases considerably with larger treatment numbers.

Estimation of genotypic and environmental variances in field nurseries (Dragavcev & Pešek 1977)

The frequency of genotypes with the desired expression of economically important quantitative characters within a hybrid population is usually very low. Therefore, the early identification and selection of such genotypes involves the analysis of very large populations. Because the breeding values of individuals in a population are masked by soil heterogeneity, environmental, competitional, and ontogenic noises, special quantitative genetic methods of analysis have to be used in order to eliminate their disturbing effects. The approach to such an analysis by using either a simple background character or an index obtained as a linear function of more background characters is described. It is believed that this approach could increase the efficiency of selection and limit the time needed to produce improved cultivars.

References

Cochran, W.G. & G.M. Cox, 1957. Experimental Designs (2nd ed.). J. Wiley and Sons, New York.

Dragavcev, V.A. & J. Pešek, 1977. Estimation of genotypic an environmental variation in plants. In: Genetic diversity in plants. Lahore, Pakistan, March 1-7, 1976, pp. 133-241.

Pešek, J., 1973. Efficiency of check plots to control the soil heterogeneity in field trials. Biom. Z. 15: 403-410.

Pešek, J., 1974. The efficiency of controls in balanced in complete block designs. Biom. Z., 16: 21-26

Smith, H.F., 1938. Am empirical rule describing heterogeneity in the yields of agricultural crops. J. Agr. Sci. 28, 1-23.

Use of oligonucleotide fingerprinting for the identification of tomato cultivars. Comparison with random amplified polymorphic DNA (RAPD)

M.J.M. Smulders, W. Rus-Kortekaas, P. Arens & B. Vosman, Centre for Plant Breeding and Reproduction Research (CPRO-DLO), P.O. Box 16, 6700 AA Wageningen, the Netherlands

Introduction

In *Lycopersicon esculentum* the genetic diversity seems to be very limited. Neither isoenzymes nor RFLP probes detect significant levels of polymorphism between tomato cultivars (Van der Beek *et al.* 1992). We have analyzed the potentials for detection of differences at the DNA level of two recently developed DNA profiling techniques: oligonucleotide fingerprinting and RAPDs. Microsatellite-containing DNA is repetitive DNA that has been found to be highly polymorphic. RAPDs are thought to detect a lower rate of polymorphism compared to microsatellites, but this has never been compared directly. The aim of this study was to compare the methods directly at two levels of variation: a low level of variation among *Lycopersicon esculentum* cultivars, and a higher level of variation among *Lycopersicon* species.

Oligonucleotide fingerprinting

When using oligonucleotide probes, such as $(GATA)_4$, $(GACA)_4$, or $(GGAT)_4$, highly polymorphic DNA regions can be detected. With $(GATA)_4$ or $(GACA)_4$ as probe, 15 tomato cultivars -including some closely related ones- could be identified by unique DNA fingerprints. A conservative analysis of the banding patterns indicated a mean band-sharing percentage of 18% for GATA-containing bands, and 51% for GACA-containing bands. The fingerprints obtained were stable during tissue culture and segregated in a Mendelian fashion (Vosman *et al.* 1992). The fingerprints of accessions of five wild *Lycopersicon* species were very different from each other, with an average band-sharing percentage between species of only 13% for GACA-containing bands.

The localization of GATA- and GACA-containing DNA fragments on the molecular map of tomato was established with an F_2-population of a cross between *L. esculentum* and *L. pennellii*. Twenty-eight loci could be mapped on 8 of the 12 tomato chromosomes. The majority of GATA- and GACA-containing loci were found to cluster (Arens *et al.* 1995) in the supposed centromeric and telomeric regions of the chromosomes.

RAPDs

As comparison, RAPDs were tested (Rus-Kortekaas *et al.* 1994). Eighty-five of 89 primers tested showed a polymorphism between *L. pennellii* and *L. esculentum*, but only four distinguished among three *L. esculentum* cultivars. These four primers could easily distinguish accessions of *Lycopersicon* species, although the average band-sharing percentage between species was relatively high (48%). When the four primers were used on the fifteen cultivars, only eleven had a unique combination of four profiles, corresponding to the possibility to distinguish 95 of the 105 possible combinations of two cultivars. In line with this, the average band-sharing percentage between cultivars was high (83%).

Comparison of the methods

The large differences between the two methods in the average percentage of bands shared, are also reflected in the fractions polymorphic and unique bands. For instance, 100% of the bands detected by GACA among the five species were polymorphic, and more than half of them were present in one accession only. In contrast, only 80% of the bands produced by the RAPD primers were polymorphic among the accessions, and of these, less than half was unique. Among the cultivars, the difference was even larger: 95% polymorphic GACA-containing bands, versus 44% polymorphic bands amplified with RAPDs.

Discussion

Microsatellite-containing DNA is repetitive DNA that has been found to be highly polymorphic. The high rate of polymorphism is thought to be due to slippage of the DNA polymerase, in combination with point mutations, unequal cross-over and recombinational events (Tautz *et al. 1986*). However, the polymorphic fragments that were detected with GATA and GACA were very large (between 1.5 and 10 kilobases)(Vosman *et al.* 1992, Arens *et al.* 1995). The causes for the polymorphism of these fragments are not clear.

RAPD bands require the presence of the primer sequence and its inverted repeat within a certain number of base pairs. Therefore, RAPDs may be generated, at least for a part, on repetitive DNA. In this study, it was shown that RAPD-generated bands show a lower rate of polymorphism compared to microsatellite-detected bands. Therefore, as far as RAPD bands are amplified from repetitive DNA, this DNA does not have the very high variation typical of microsatellite-containing DNA.

Conclusions

Taken together, the results presented here indicate that the two methods detect DNA with a different degree of variation. This has implications for the potential use of the two methods.

GATA- or GACA-fingerprinting appears a very powerful method to distinguish among genetically very related material, such as modern tomato cultivars. It is not yet clear how useful these microsatellite sequences are for the determination of the relationships between less closely related plants, since they may have no band in common, as in the case of the bands detected by $(GATA)_4$ among accessions of different *Lycopersicon* species. Perhaps, less polymorphic microsatellites are necessary to study this kind of material.

RAPDs do not detect enough variation to easily distinguish all tomato cultivars, but they do detect fragments of the same size in accessions of different *Lycopersicon* species. Provided that this method proves to be easily reproducible, and that the fragments of the same size do represent identical or closely related sequences, RAPDs may be suited for studies of genetic diversity among *Lycopersicon* species.

References

Arens, P., P. Odinot, A.W. van Heusden, P. Lindhout & B. Vosman, 1995. Genome: submitted.
Rus-Kortekaas, W., M.J.M. Smulders, P. Arens & B. Vosman, 1994. Genome 37: 375-381.
Tautz, D., M. Trick & G.A. Dover, 1986. Nature 322: 652-656.
Van der Beek, H., R. Verkerk, P. Zabel & P. Lindhout, 1992. Theor. Appl. Genet. 84: 106-112.
Vosman, B., P. Arens, W. Rus-Kortekaas & M.J.M. Smulders, 1992. Theor. Appl. Genet. 85: 239-244

Genetics and mapping of new isozyme loci in *Vicia faba* L. using trisomics

A.M. Torres[1], Z. Satovic[2], J. Canovas[1], S. Cobos[1] & J.I. Cubero[3], [1]C.I.D.A., Departamento de Mejora y Agronoma, Aptdo. 4240, 14080 Córdoba, Spain, [2]Dept. of Plant Breeding, Genetics and Biometrics, Faculty of Agriculture, 41000 Zagreb, Croatia, [3]E.T.S.I.A.M., Departamento de Genética, Aptdo. 3048, 14080 Córdoba, Spain.

Key words
gene mapping, isozymes, trisomic, *Vicia faba*

In comparison with other pulse crops, faba bean has been the subject of relatively little research and only few isozyme markers have been studied. Our goal is to study the genetics of newly detected loci and to add new isozyme markers to the preliminary map of the species.

The polymorphism in ten enzyme systems in faba bean (*Vicia faba* L.) is being analyzed using horizontal starch gel electrophoresis (Gotlieb 1973). The study has revealed thirteen loci most of which have not been reported before (Gates & Boulter 1979, Mancini *et al.* 1989, Suso & Moreno 1982, Suso & Moreno 1983, Peat & Adham 1984, Torres *et al.* 1993). The systems assayed (Wendel & Weeden 1990) include aspartate aminotransferase (AAT), aconitase (ACO), acid phosfatase (ACP), esterase (EST), fructose kinase (FK), malic enzyme (ME), nacyl glucose aminidase (NAG), peroxidase (PRX), 6-phosphogluconate dehydrogenase (6PGD), and superoxide dismutase (SOD).

Segregation and linkage analysis are being performed using the computer programs Linkage-1 (Suiter *et al.* 1983) and MAPMAKER (Lander *et al.* 1987). Each of the thirteen loci exhibits monogenic inheritance and most of them have shown independent assortment, probably due to the low number of isozymes studied so far. Chromosomal location of isozyme loci has been determined based on deviating F2 segregation of plants trisomic for four of the six chromosomes of *V. faba* (III, IV, V and VI) (Martín & Barceló 1984). Primary trisomics provide an excellent cytogenetic tool to assign genes to specific chromosomes. The normal codominant ratio 1:2:1 is expected to be found in all F2 disomic populations, except those involving a third chromosome carrying the isozyme to be located. In this case the ratio is modified because of the presence of the extra chromosome.

In the present study, five loci have been unambiguously assigned to a specific chromosome: Est-2 to chromosome III, Fk-1 to chromosome IV, Prx-1 to chromosome V and Sod-1 and Pgd-1 to chromosome VI (Hermsen 1970). The typical Mendelian segregation observed for the rest of the loci also clearly indicates that they are not located on the chromosomes considered in the study. The fact that most of the markers analyzed so far seem to segregate independently (except for Nag-1 and Pgd-2) is not at all surprising. If we consider the enormous size of the *V. faba* genome, linkage between such a small number of studied characters is unlikely. The study of additional allozymic variants in wider faba bean crosses and the inclusion of RAPD and RFLP markers in our analysis will allow to identify new linkage groups that so far have escaped gene mapping.

References

Gates, P. & D. Boulter, 1979. The use of seed isozymes as an aid to the breeding of field beans (*Vicia faba* L.). New Phytol. 84: 501-504.

Gotlieb, L.D., 1973. Enzyme differentiation and phylogeny in *Clarkia franciscana, C. rubicunda* and *C. amoena*. Evolution 27: 205-214.

Hermsen, H.G.Th., 1970. Basic information for the use of primary trisomics in genetic and breeding research. Euphytica 19: 125-140.

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

Mancini, R., C. De Pace, G.T. Scarascia, V. Delre & D. Vitori, 1989. Isozyme gene markers in *Vicia faba* L. Theor. Appl. Genet 77: 657-667.

Martín, A. & P. Barceló, 1984. The cytology and morphology of Vicia faba trisomics. In: G.P. Chapman & S.A. Tarawai (Eds.), pp. 63-76. Systems for cytogenetic analysis in *Vicia faba* L. Martinus Nijhoff, the Hague.

Peat, W.E. & J.Y. Adham, 1984. The use of isozyme genes as markers in the population genetics of *Vicia faba* L. In: G.P. Chapman & S.A. Tarawali (Eds.). Systems for cytogenetic analysis in *Vicia faba*, Advances in Agricultural Biotechnology, Vol. II, pp. 109-117. M. Nijhoff, Dr. W. Jaak Publishers, Dordrecht, the Netherlands.

Suiter, K., J. Wendel & J. Case, 1983. LINKAGE-1: a Pascal computer program for the detection and analysis of genetic linkage. J. Hered. 74: 203-204.

Suso, M.J. & M.T. Moreno, 1982. Genetic control of electrophoretic variants of Glutamate oxalacetate transaminase (GOT) in *Vicia faba* L. FABIS 5, 14.

Suso, M.J. & M.T. Moreno, 1983. Isozymatic polymorphism for superoxide dismutase (SOD) in *Vicia faba* and its systematic implications. FABIS 16, 3-5.

Torres, A.M., N.F. Weeden and A. Martín, 1993. Linkage among isozyme, RFLP and RAPD markers in *Vicia faba*. Theor. Appl. Genet. 85: 937-945.

Wendel, J.F. and N.F. Weeden, 1990. Visualization and interpretation of plant isozymes. In: D.E. Soltis and P.S. Soltis (Eds.). Isozymes in plant biology, pp. 5-45. Dioscorides Press, Portland, Oregon.

Faba bean mapping with trisomics. Linkage among morphological, isozyme, and RAPD markers

A.M. Torres[1], Z. Satovic[2], J. Canovas[1], A. Martín[3] & J.I. Cubero[4], [1]C.I.D.A., Departamento de Mejora y Agronomía, Apdo. 4240, 14080 Córdoba, Spain, [2]Dept. of Plant Breeding, Genetics, and Biometrics, Faculty of Agriculture, 41000 Zagreb, Croatia, [3]C.S.I.C.-I.A.S., Apdo 4084, 14080 Córdoba, Spain, [4]E.T.S.I.A.M., Departamento de Genética, Apdo. 3048, 14080 Córdoba, Spain

Key words

isozymes, RAPD markers, linkage groups, trisomic, *Vicia faba*

The construction of linkage maps greatly increases the efficiency of genetic and breeding studies. In comparison with other legumes such as garden pea, the faba bean (*Vicia faba* L.) has been the subject of little research in this respect, and up to now only a few extended linkage groups have been described (Torres *et al.* 1993). The number of cytological tools available in faba bean, for assigning genes and linkage groups to their respective chromosomes, is limited to translocation stocks (Sjödin 1971) and primary trisomics (Cabrera & Martín 1989, Cabrera *et al.* 1989). To date, five of a possible six primary trisomics have been characterized by our group (Martín & Barceló 1984). This offers a useful tool to enhance the preliminary map of this species.

In this study, chromosomal locations and linkages among several morphological, isozyme, and RAPD markers are being investigated by using the primary trisomics III, IV, V and VI. Thirteen F2 populations derived from these trisomics are being scored for morphological and allozyme phenotypes, and seven of them are also being analyzed for RAPD polymorphisms. Morphological traits studied so far include: determinate growth (Ti/ti), unifoliate (Una^1/una^1), red seed-coat (R/r), solid distribution of pigment on flower (Sdp/sdp), yellow pigment on flower (Yf/yf), hylum colour (N/n), and anthocyanin content in stem. The following enzyme systems are being considered: aspartate aminotransferase (AAT), aconitase (ACO), esterase (EST), fructose kinase (FK), malic enzyme (ME), mannose phosphate isomerase (MPI), nacyl glucose aminidase (NAG), peroxidase (PRX), 6-phosphogluconate dehydrogenase (6PGD), superoxide dismutase (SOD) and triose phosphate isomerase (TPI) (Wendel & Weeden 1990). With regard to RAPD markers, a total of 98 oligonucleotide primers have been surveyed in the parental lines involved in the crosses. Each primer yields between one to eight scorable loci. The segregation of the "present" and "absent" phenotypes usually provides a good fit to the expected 3:1 ratio.

Goodness-of-fit to the expected F2 segregations are tested by Chi-square analysis. Linkage among all considered markers is studied from F_2 segregations using maximum-likelihood formulae by the programs Linkage-1 (Suiter *et al.* 1983) and MAPMAKER (Lander *et al.* 1987). The normal codominant ratio 1:2:1 for isozymes or 3:1 for morphological and RAPD markers is expected in all F2 disomic populations. In the populations involving a third chromosome carrying the marker to be located, the previously mentioned ratios are modified because of the presence of the extra chromosome (Martín & Barceló 1984). Both positive and negative results on linkage and chromosomal location will provide new information useful for the construction of a more complete map of faba bean.

At present, three of the seven crosses have been analyzed and their data have been studied separately. Crosses 6x2 TV and 6x33 TIV displayed eighteen linkage groups each, and cross 6x159 revealed sixteen. Although more linkage groups have been identified than there are chromosomes, at present we are unable to determine which linkage groups are syntenic, or whether we have markers on every chromosome. After analyzing the remaining crosses, data will be pooled to map partially overlapping sets of informative genetic markers. Homogeneity of recombination among populations will be tested, and a composite linkage map based upon the seven F2 populations will be presented.

References

Cabrera, A. & A. Martín, 1989. Analysis of genetic linkage in faba bean (*Vicia faba* L.). Fabis 24: 3-5.

Cabrera, A., J.I. Cubero & A. Martín (1989). Genetic mapping using trisomics in *Vicia faba* L. Fabis 23: 5-7.

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

Martín, A. and P. Barceló, 1984. The cytology and morphology of *Vicia faba* trisomics. In: G.P. Chapman and S.A. Tarawai (Eds.). Systems for cytogenetic analysis in *Vicia faba* L., pp. 63-76. Martinus Nijhoff, the Hague.

Sjödin, J., 1971. Induced morphological variation in *Vicia faba* L.. Hereditas 67: 155-180.

Suiter, K., J. Wendel & J. Case, 1983. LINKAGE-1: a Pascal computer program for the detection and analysis of genetic linkage. J. Hered. 74: 203-204.

Torres, A.M., N.F. Weeden & A. Martín, 1993. Linkage among isozyme, RFLP and RAPD markers in *Vicia faba*. Theor. Appl. Genet. 85: 937-945.

Wendel, J.F. & N.F. Weeden, 1990. Visualization and interpretation of plant isozymes. In: D.E. Soltis and P.S. Soltis (Eds.). Isozymes in plant biology, pp. 5-45. Dioscorides Press, Portland, Oregon.

Multiple alleles detected of a QTL for tuber shape in potato

Herman J. van Eck[1], *Jeanne M.E. Jacobs*[2], *W.J. Stiekema*[2] *& E. Jacobsen*[1], [1]*Department of Plant Breeding, Wageningen Agricultural University, P.O. Box 386, 6700 AJ Wageningen,* [2]*DLO-Centre for Plant Breeding and Reproduction Research (CPRO-DLO), Department of Molecular Biology, P.O. Box 16, 6700 AA Wageningen, the Netherlands*

Introduction

This paper describes the localization of a quantitative trait locus (QTL) in the F_1 progeny from non-inbred parents, *i.e.*, a QTL for tuber shape in potato. The shape of potato tubers is analyzed quantitatively as well as qualitatively. Tuber shape is commonly regarded as a quantitative character because of the continuous variation ranging from round, via oval to long. However, among the clones of a diploid full-sib population, it is possible to discern visually between two distinct phenotypic classes: round and long.

Results

Full details can be found in Van Eck *et al.* (1994). On the basis of this visual classification the inheritance of tuber shape is explained by presuming a monogenic dominant locus Ro, round being dominant to long. Both parents were heterozygous round (Ro ro), and the observed segregation 68:29 fits a 3:1 ratio. With RFLPs the Ro locus was mapped on chromosome 10 using normal linkage analysis. Tuber shape was also studied as a quantitative trait, using the length/width ratio as phenotypic value. The broad sense heritability, based on variation between clones and between tubers within clones, was equal to 0.80. The morphologically mapped Ro locus could explain 75% of the genetic variation, indicating the presence of a major QTL at the Ro locus and minor quantitative genetic factors outside of it.

The linkage phase of alleles from adjacent loci can be determined on the basis of cosegregation. By using this type of information about linkage between unique alleles of flanking RFLPs in coupling phase with Ro alleles, it was possible to identify the origin of the alleles at the Ro locus. The 3:1 (round:long) segregating progeny was divided into four genotypic classes specified by their allelic composition: $\mathrm{Ro}^{♀}\mathrm{Ro}^{♂}$: $\mathrm{Ro}^{♀}\mathrm{ro}$: $\mathrm{ro}\,\mathrm{Ro}^{♂}$:

ro ro = 1:1:1:1. The recessive ro allele is identical by descent in both parents. The effect on tuber shape of the non-identical alleles Ro$^{♀}$, Ro$^{♂}$ en ro was evaluated by comparing the mean length/width ratio of the four genotypic classes. The heterozygous genotypes Ro$^{♀}$ro and ro Ro$^{♂}$ differed significantly in their length/width ratio ($p = 0.016$). This difference in length/width ratio was explained by postulating multiple alleles at the Ro locus.

Discussion

From the presence of multiple alleles it is conceivable why at the tetraploid level never a monogenic inheritance for tuber shape was described. Complex intralocus interactions between multiple Ro alleles cause at the tetraploid level a continuous variation for tuber shape. In fact, in diploid potato multi-allelism is observed at approximately one third of the RFLP loci. Therefore, multi-allelism for QTLs may be an underestimated source of quantitative genetic variation. Accordingly, it would be reasonable to assume that also for other quantitative traits the hereditary basis is determined by multiple loci in combination with multiple alleles.

A procedure to localize the positions of quantitative trait loci in the offspring of heterozygous parents ought to take into account the above explained situations. The difference between the classes which are indicated by the molecular markers can be annulled by intralocus interactions of QTL-alleles. Usage of dominant markers like Random Amplified Polymorphic DNA markers (RAPDs) as well as the use of codominant markers which are heterozygous in only one of the parents should be avoided in QTL-mapping experiments. These type of markers can identify only two out of the four possible classes. Multi-allelic RFLPs are the most appropriate type of markers since they can discriminate between the four possible allele combinations in the offspring of non-inbred parents, allowing unbiased estimation of the presence and the magnitude of QTLs.

Reference

Eck, H.J. van, J.M.E. Jacobs, P. Stam, J. Ton, W.J. Stiekema & E. Jacobsen, 1994. Multiple alleles for tuber shape in diploid potato detected by qualitative and quantitative genetic analysis using RFLPs. Genetics 137: 303-309.

Insect resistance and molecular markers in chrysanthemum analysed with a newly developed regression (WeSel)

K. Wolff[1,2]*, A.F.M. Nierop*[3] *& C.M. de Jager*[1]*,* [1]*University of Leiden, Institute of Evolutionary and Ecological Sciences, PO Box 9516, 2300 RA Leiden,* [2]*TNO Nutrition and Food Research, Dept. of Microbiology,* [3]*TNO Nutrition and Food Research, Centre for Structure Elucidation and Instrumental Analysis, PO Box 360, 3700 AJ Zeist, The Netherlands*

Chrysanthemum has been bred for over 3000 years, and nowadays it is a very popular cut flower with a high economic value. The aim of the present study is to find and characterize genetic markers in order to localise genes involved in insect resistance.

Chrysanthemum is part of a hexaploid species complex, and has a chromosome number of around 54 chromosomes. Because of the strong self-incompatibility system, no inbred lines are available. The offspring of a biparental cross (F_1) demonstrates a wide range of morphological characters, and can be used as the segregating family.

The traits analysed are the number of leaf miner pupae and two types of thrips feeding damage (silver- and growth-damage). Sixty offspring of a test cross were analysed. Each plant was tested in fivefold for leafminer damage in a non-choice experiment, and in eightfold for thrips damage in a choice experiment (De Jager *et al.* 1995). The genetic markers used are RAPDs and RFLPs (Wolff *et al.* 1993, 1994). RAPDs are dominant markers by nature. The autoradiograms showed multiple fragments per lane (6-12), and, therefore, these fragments also behave as dominant characters. Consequently, only the presence or absence of each RAPD or RFLP fragment was noted.

A total of 384 polymorphic markers were scored in order to have several markers on each of the chromosomes of both parents. A novel method of regression analysis with weighted selection (WeSel) was developed for an optimal prediction of insect resistance traits. The prediction is optimal if it is as stable and accurate as possible, and if a minimum number of markers is selected as predictor set.

Significant differences between the offspring for the insect resistance traits were found. A preliminary analysis showed that several RFLP and RAPD markers have a significant correlation with the insect resistance characters (Table 1). WeSel was used to test which markers, in combination with each other, optimally predict the resistance found (Table 2). Seven markers are sufficient to explain 49% of silver damage, three markers predict 52% of growth damage, and for the number of pupae three markers explain 51% of the variation. Note that of the four markers in Table 2, selected for a

Table 1. RAPD and RFLP markers that show a highly significant association ($P < 0.01$) with insect resistance characters, their significance level and the % of variance explained by each marker

Silver damage			Growth damage			Nr. of pupae		
Marker	P	%var	Marker	P	%var	Marker	P	%var
RAPD-A16.3	0.003	15.7	RAPD-22.5	0.000	37.8	RAPD-4.2	0.000	25.3
RFLP-364.7	0.006	13.2	RAPD-26.3	0.002	20.4	RAPD-10.1	0.004	18.0
			RAPD-27.7	0.004	15.0	RAPD-10.2	0.004	18.2
			RFLP-263.6	0.006	12.9	RAPD-11.5	0.004	12.7
						RAPD-27.1	0.001	17.4
						RAPD-33.5	0.001	19.0
						RFLP-423.3	0.006	13.0
						RFLP-391.4	0.007	11.9

Table 2. Markers that, acting together, give the best prediction of insect resistance traits, as analysed with the WeSel program

Trait	RAPD markers	RFLP markers
Silver damage	11.5 31.1 A4.13 A16.3 A16.4	364.7 450.5
Growth damage	22.5 27.7	263.6
Nr. of pupae	B20.3	432.2 263.2

highly significant prediction of the number of pupae, only one is among the highly significant markers in Table 1.

The segregation of the markers showed that chrysanthemum is mainly an autopolyploid by origin, as indicated by their segregation ratios and linkage of fragments in coupling.

Our goal now is to sequence the important RAPD and RFLP markers and develop specific primer sets. These primers will first be tested in the present test cross. Whether successful primer sets are transferable to other test crosses and cultivars is the next important question to answer.

Acknowledgements

This is a publication of the PBTS-project "Rapid biotechnological detection of resistance in chrysanthemum against insect pests", which is a joint activity of the University of Leiden, TNO-Zeist, and the chrysanthemum breeders CBA, Fides, and Hoek Breeding. It is subsidized by Stipt (Ministry of Economic Affairs).

References

De Jager, C.M., R.P.T. Butôt, P.G.L. Klinkhamer, T.J. De Jong, K. Wolff & E. Van Der Meijden, 1995. Genetic variation in chrysanthemum for resistance to thrips. In preparation.

Wolff, K. & J. Peters-Van Rijn, 1993. Rapid detection of genetic variability in chrysanthemum (*Dendranthema grandiflora* Tzvelev) using random primers. Heredity 71: 335-341.

Wolff, K., J. Peters-Van Rijn & H. Hofstra, 1994. RFLP analysis in chrysanthemum. I. Probe and primer development. Theor. Appl. Genet.: in press.

Efficiency of single root selection in a full-sib family breeding programme in sugar beet

B. Zhao, I.J. Mackay, P.D.S. Caligari & R. Mead, Departments of Agricultural Botany and Applied Statistics, The University of Reading, and Lion Seeds Ltd., Maldon, Essex, U.K.

Introduction

Full-sib family recurrent selection has been demonstrated to be an important means of improving characters in sugar beet for which there are little or no heterotic effects (Bosemark 1993, Hecker & Helmerick 1985). It has been one of the main methods used in the improvement of source populations at Lion Seeds Ltd. In this programme, full-sib families are selected on the basis of results from replicated field trials. Individual roots within each selected full-sib family are then selected from the nursery plots on the basis of root shape, root weight, sugar content and juice purity. The following season, these roots are planted and the resulting plants are crossed in all possible pairs to produce full-sib families for the next generation of selection. Many suitable parental lines have been extracted from populations continuously improved by this method and have been used successfully in hybrid production.

However, information is lacking regarding the contribution of both full-sib family selection and the above mentioned within family single root selection to the effectiveness of the population improvement programme. The objective of this study is to examine the efficiency of within family single root selection as well as the efficiency of full-sib family selection.

Materials and Methods

Historical data from two cycles of full-sib family recurrent selection on a multigerm population have been studied; the first cycle from 1988 to 1990, the second from 1990 to 1992. Correlation coefficients were estimated (1) between mid-parent and offspring and (2) between the selected full-sib families in the parental generation and GCA effects estimated on the progeny in the next generation.

The mid-parent data were obtained from single roots selected in the nursery. The

GCA effects of the selected full-sib families were estimated using Griffings method 4, model 2 (Griffing 1956) since they had been selected from the previous cycle. For ease of estimation, using GENSTAT, the double copy method (Thompson 1984) has been used. This method uses two copies of the data prepared in such a way that male and female parents in the first copy will be the female and male parents in the second copy. In the analysis of variance, variety (full-sib family) effects are partitioned into three components, male parent effect, female parent effect (both of which can be thought as the GCA effects and which should be identical), and the male × female interaction which is essentially the SCA.

Results

Similar results have been obtained for the two rounds of selections. Correlations between mid-parent and offspring are high for sugar content and juice purity characters, but low for root weight for both data sets. The second set of correlations, between selected full-sib families in the parental generation and the GCA effects estimated on the progeny, are still high for sugar content and juice purity. For root weight, correlations are higher than those between mid-parent and offspring, but remain relatively low.

The pattern of the correlation coefficients found is consistent with both full-sib family selection and single plant selection being effective in the improvement of sugar content and juice purity in sugar beet. Single plant selection is not effective in improving root weight. This character should, however, show some response to full-sib family selection.

It is problematical with such historical data to quantify the relative contributions of between and within family selection to the total response to selection. Further work will address this question by collecting data from designed selection experiments. The selection programme can then be optimised to get the maximum benefit from the between and within family components.

Table 1. Correlation coefficients across generations

	1st set		2nd set	
	88/90	90/92	88/90	90/92
Sugar Content	0.31	0.35	0.59	0.58
Root Weight	0.16	0.15	0.43	0.35
K (100/w)	0.48	0.52	0.22	0.44
Na (100/w)	0.45	0.49	0.65	0.68
NH2 (100/w)	0.38	0.45	0.61	0.70

Acknowledgement

This work has been made possible by financial support from Lion Seeds Ltd., Maldon, Essex, UK.

References

Bosemark, N.O., 1993. Genetics and breeding. In: D.A. Cooke & R.K. Scott (Ed.), The Sugar Beet Crop. Chapman & Hall.

Griffing, B., 1956. A generalized treatment of the use of diallel crosses in quantitative inheritance. Heredity 10: 31-50.

Hecker, R.J. & R.H. Helmerick, 1985. Sugar beet breeding in the United States. In: G.E. Russell (Ed.), Progress in Plant Breeding, vol.1. Butterworths, London.

Thompson, R., 1984. The use of multiple copies of data in forming and interpreting analysis of variance. In: K. Hinkelmann (Ed.), Experimental design, statistical model and genetic statistics. Marcel Dekker, New York.

List of participants

Austria

Buerstmayr, H.	Institute of Plant Breeding, University of Agriculture, Gregor Händelstr. 33, 1180 Vienna

Belgium

Botterman, J.	Plant Genetic Systems NV, Jozef Plateaustraat 22, 9000 Gent
Laecke, K. van	Rijksstation voor Plantenveredeling, RVP, B. van Gansberghelaan 109, 9820 Merelbeke
Loose, M. de	Rijksstation voor Plantenveredeling, RVP, B. van Gansberghelaan 109, 9820 Merelbeke

Croatia

Pecina, M.	University of Zagreb, Faculty of Agriculture, Svetosimunska 25, 41000 Zagreb
Vasilj, D.	University of Zagreb, Faculty of Agriculture, Svetosimunska 25, 41000 Zagreb

Czech Republic

Pešek, J.	RI of Animal Nutrition Pohorelice, Department of Soil Management, 66462 Hrusovany near Brno

Denmark

Kjær, B.	RISO National Laboratory, Plant Biology Section, P.O.Box 49, 4000 Roskilde
Pedersen, C.A.	DLF-Trifolium, P.O.Box 19, 4660 Store Heddinge

Finland

Tammisola, J.M.	VTT Biotechnology and Food Research, P.O.Box 1505, 02044 VTT Espoo

France

Bar-Hen, A.	GEVES La Minière, 78280 Guyancourt cedex
Charcosset, A.	INRA, Ferme du Moulon, 91190 Gif-s/Yvette
Cilas, C.A.J.	CIRAD-CP, U.R. Biometric, P.O.Box 5035, 34032 Montpellier cedex
Combat, B.	Biofords Consultants, 8 Ave de Montespan, 91024 Evry cedex
Fatmi, A.	Biofords Consultants, 8 Ave de Montespan, 91024 Evry cedex
Flori, A.	CIRAD-CP, U.R. Biometric, P.O.Box 5035, 34032 Montpellier cedex
Gallais, A.	INRA-UPS, Station de Génétique Végétale, Ferme du Moulon, 91190 Gif-s/Yvette
Goldringer, I.	Station de Génétique Végétale, Ferme du Moulon, 91190 Gif-s/Yvette
Hospital, F.	INRA, Station de Génétique Végétale, Ferme du Moulon, 91190 Gif-s/Yvette
Karaman, Z.	Limagrain Genetics, P.O.Box 115, 63203 Riom cedex
Lorieux, M.J.	AGETROP-CIRAD, P.O.Box 5035, 34032 Montpellier cedex
Mangin, B.	INRA, Laboratoire de Biometrie et d'Intelligence Artificielle, P.O.Box 27, 31326 Castanet-Tolosan
Rebaï, A.	INRA, Laboratoire de Biometrie et d'Intelligence Artificielle, P.O.Box 27, 31326 Castanet-Tolosan

Germany

BOHN, M. Universität Hohenheim, Institut 350, 70593 Stuttgart

BORCHARDT, D. Universität Hohenheim, Institut für Pflanzenzüchtung, Saatgutforschung und Pop.genetik (350), 70593 Stuttgart

ENGEL, F. A. Dieckmann-Heimburg, Saatzucht Sulbeck, P.O.Box 1165, 31684 Nienstadt

FAHR, S. Universität Hohenheim, Institut für Pflanzenzüchtung, Saatgutforschung und Pop.genetik (350), 70593 Stuttgart

GEIGER, H.H. Universität Hohenheim, Institut für Pflanzenzüchtung, Saatgutforschung und Pop.genetik (350), 70593 Stuttgart

HAUSSMANN, B.I.G. Universität Hohenheim, Institut für Pflanzenzüchtung, Saatgutforschung und Pop.genetik (350), 70593 Stuttgart

JANSEN, R. KWS Kleinwanzlebener Saatzucht AG, P.O.Box 1463, 37555 Einbeck

LÉON, J. University of Kiel, Inst. f. Pflanzenbau und Planzenzüchtung, 24098 Kiel

LOOCK, A. KWS Kleinwanzlebener Saatzucht AG, P.O.Box 1463, 37555 Einbeck

MELCHINGER, A.E. University of Hohenheim, 350 Institut für Pflanzenzüchtung, Saatgutforschung und Populationsgenetik, 70593 Stuttgart

MÜLLER, P. Institut für Angewandte Genetik, Herrenhauserstrasse 2, 30419 Hannover

PIEPHO, H. University of Kassel, FB 11, Steinstrasse 19, 37213 Witzenhausen

SCHIPPRACK, W. Fr. Strube Saatzucht Kg, Maizuchtstation Süd, Der Hohe weg zum Rhein 16, 68307 Mannheim

SCHNELL, F.W. Universität Hohenheim, 350 Institut für Pflanzenzüchtung, Saatgutforschung und Populationsgenetik, 70593 Stuttgart

SCHÖN, C. KWS Kleinwanzlebener Saatzucht AG, P.O.Box 1463, 37555 Einbeck

STEINRUCKEN, G. A. Dieckmann-Heimburg Saatzucht Sulbeck, P.O.Box 1165, 31684 Nienstadt

UTZ, F. University of Hohenheim, 350 Institut für Pflanzenzüchtung, Saatgutforschung und Populationsgenetik, 70593 Stuttgart

WEBER, E. Inst. für Pflanzenzüchtung & Saatgutwirtschaft, Martin-Luther-Univ. Halle, Wittenberg. Berliner Strasse 2, 06188 Hohenthurm

Hungary

HAJOS-NOVAK, M. Gödöllő Agricultural University, Dept. of Genetics and Plant Breeding, P.O.Box 303, 2103 Gödöllő

Italy

CAMUSSI, A. Instituto di Selvicoltura Cattedra di Genetica Agraria, Via S. Bonaventura 13, 50145 Firenze

NOVARO, P. Experimental Inst. for Cereal Research, Via Cassia 176, 00191 Rome

Malaysia

CHEAH, S.C. Palm Oil Research Intitute of Malaysia, P.O.Box 10620, 50720 Kuala Lumpur

Netherlands

BEEREPOOT, L.J. Barenbrug Holland B.V., Postbus 4, 6678 ZG Oosterhout

BOS, I. Wageningen Agricultural University, Dept. of Plant Breeding, Postbus 386, 6700 AJ Wageningen

CAPPELLEN, W. VAN Bejo Zaden BV, Postbus 50, 1749 ZH Warmenhuizen

DOURLEIJN, J. Wageningen Agricultural University, Dreijenlaan 4, 6703 HA Wageningen

ECK, H.J. VAN Wageningen Agricultural University, Dept. of Plant Breeding, Postbus 386, 6700 AJ Wageningen

EEUWIJK, F.A. VAN CPRO-DLO, Postbus 16, 6700 AA Wageningen

EGGERMOND, A. VAN De Ruiterzonen, Postbus 4, 2665 ZG Bleiswijk

Geling, K.B.	Royal Vanderhave Group, P.O. Box 1, 4420 AA Kapelle
Ghijsen, H.C.H.	CPRO-DLO, Postbus 16, 6700 AA Wageningen
Heijden van der, S.	CEBECO Zaden BV, Postbus 10000, 5250 GA Vlijmen
Jacobs, J.M.E.	CPRO-DLO, Postbus 16, 6700 AA Wageningen
Jansen, J.	CPRO-DLO, Postbus 16, 6700 AA Wageningen
Jansen, R.C.	CPRO-DLO, Postbus 16, 6700 AA Wageningen
Keizer, L.C.P.	CPRO-DLO, Postbus 16, 6700 AA Wageningen
Knaap, J.C.M. van der	Fides Research and Breeding BV, Postbus 26, 2678 ZG De Lier
Kraakman, A.T.W.	Wageningen Agricultural University, Dept. of Plant Breeding, Postbus 386, 6700 AJ Wageningen
Kramer, T.	Asgrow/Bruinsma Seeds, Postbus 24, 2670 AA Naaldwijk
Kuiper, M.	Keygene N.V., Postbus 216, 6700 AE Wageningen
Lindhout, P.	Wageningen Agricultural University, Dept. of Plant Breeding, Postbus 386, 6700 AJ Wageningen
Maliepaard, C.H.	CPRO-DLO, Postbus 16, 6700 AA Wageningen
Morales, M.	Van der Have Research, Postbus 1, 4410 AA Rilland
Neele, A.E.F.	CEBECO Zaden BV, Postbus 10000, 5250 GA Vlijmen
Nijs, T.P.M. den	CPRO-DLO, Postbus 16, 6700 AA Wageningen
Oeveren, A.J. van	Thorbeckestraat 42, 6702 BS Wageningen
Ooijen, J.W. van	CPRO-DLO, Postbus 16, 6700 AA Wageningen
Oostveen, B.C.M.	Nunhems Zaden BV, Postbus 4005, 6080 AA Haelen
Otten - Van der Velde, A.J.	Florigene Europe BV, Waardlaan 4A, 2231 NA Rijnsburg
Peerbolte, R.	D.J. van der Have B.V., Postbus 1, 4410 AA Rilland
Peleman, J.	Keygene N.V., Postbus 216, 6700 AE Wageningen
Reinink, K.	CPRO-DLO, Postbus 16, 6700 AA Wageningen
Schut, J.W.	Wageningen Agricultural University, Dept. of Plant Breeding, Postbus 386, 6700 AJ Wageningen
Smulders, M.	CPRO-DLO, Postbus 16, 6700 AA Wageningen
Stam, P.	CPRO-DLO, Postbus 16, 6700 AA Wageningen
Vermeer, H.	CEBECO Zaden B.V., Lisdoddeweg 36, Postbus 139, 8200 AC Lelystad
Vogelaar, A.	Rijk Zwaan Zaadteelt en Zaadhandel BV, Postbus 40, 2678 ZG De Lier
Voorrips, R.E.	CPRO-DLO, Postbus 16, 6700 AA Wageningen
Wolff, K.	TNO-Zeist, Dept. of Microbiology, Postbus 360, 3700 AJ Zeist
Wolters, P.	CPRO-DLO, Postbus 16, 6700 AA Wageningen
Wolters, T.	Nunhems Zaden, Postbus 4005, 6080 AA Haelen
Woude, K. van der	Van der Have, Postbus 1, 4481 DD Kloetinge

Norway

Honne, B.I.	SFL Kvithamar, 7500 Stjoerdal

Poland

Adamski, T.	Institute of Plant Genetics, Strzeszynska 34, 60-479 Poznan
Dobek, A.	Poznan Agricultural University, Dept. of Math. & Stat. Methods, Wojska Polskiego 28, 60-637 Poznan
Kaczmarek, Z.S.	Institute of Plant Genetics, Strzeszynska 34, 60-479 Poznan
Kala, R.	Poznan Agricultural University, Dept. of Math. & Stat. Methods, Wojska Polskiego 28, 60-637 Poznan
Krajewski, P.	Institute of Plant Genetics, Strzeszynska 34, 60-479 Poznan
Łuczkiewicz, T.K.	University of Agriculture, Dept. of Genetics and Plant Breeding, Wojska Polskiego 71c, 60-625 Poznan
Surma, M.	Institute of Plant Genetics, Strzeszynska 34, 60-479 Poznan
Wasilewicz-Flis, J.	Potato Research Institute, Mtochow, 05-832 Ruzalin

Slovenia

Kralj, D. — Institute for Hop Research and Brewing, 63310 Zalec

Spain

Millan, T. — ETSIAM, Dpto Genetica, Avda Menendez Pidal S/N, Apdo 3048, 14080 Cordoba

Torres, A.M. — CIDA, Dpto Mejora y Agronomia, Alameda del Obispo S/N, Apartado 4240, 14080 Cordoba

Sweden

Becker, H. — Swedish University of Agric. Sciences, Dept. of Plant Breeding, 268 31 Svalov

Engqvist, G. — Swedish University of Agric. Sciences, Dept. of Plant Breeding, 268 31 Svalov

Johansson, E. — Hilleshoeg AB, P.O.Box 302, 26123 Landskrona

Switzerland

Gianfranceschi, L. — Swiss Fed. Inst. of Technology, Phytomedizin/Pathologie ETH/Zentrum, Universitätstrasse 2, 8092 Zürich

Ragot, M. — Ciba Seeds, R-1096 2 32, 4002 Basel

United Kingdom

Camlin, M.S. — Dept. of Agriculture for Northern Ireland, Plant Testing Station, Crossnacreevy, Belfast BT6 9SH

Chasalow, S.D. — Scottish Agricultural Statistics Service, Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA

Curnow, R.N. — University of Reading, Dept. of Applied Statistics, P.O.Box 240, Reading RG6 2FN

Graham, J. — Scottish Crop Research Institute, Invergowrie, Dundee, Scotland

Hackett, C..A. — Scottish Agricultural Statistics Service, Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA

Hayward, M.D. — Inst. of Grassland & Environm. Research, Welsh Plant Breeding Station, Aberystwyth, Wales

Hill, J. — Broneirian, Llanbadarn Road, Aberystwyth, Dyfed SY23 1HB, Wales

Hossain, K.G. — IGER, Pl Genetic and Breeding Department, Plas Gogerddan, Aberystwyth SY23 3EB

Lanham, P. — Scottish Crop Research Institute, Soft Fruit Genetics Department, S.C.R.I., Invergowrie, Dundee DD2 5DA

Mackay, I.J. — Lion Seeds Ltd., Woodham Mortimer, Maldon, Essex CM9 6SN

Pike, D. — Zeneca Seeds, Jealott's Hill Research Station, Bracknell, Berkshire RG12 6EY

Roberts, A.M.I. — Zeneca Seeds, Jealott's Hill Research Station, Bracknell, Berkshire RG12 6EY

Talbot, M. — SASS, University of Edinburgh, JCMB Kings Building, EH93J2 Edinburgh

Zhao Baidong — University of Reading, Department of Applied Statistics, P.O.Box 240, Reading RG6 2AL

U.S.A.

Brown, S. — Genetic Consultant, 2993 Pinch Road, Manheim, PA 17545-9464

Walton, M. — Linkage Genetics, 1515 West 2200 South, Suite C, Salt Lake City, Utah 84119