Comparing Molecular Marker Combinations in Natural or Artificial Populations

Vivi N. Arief
The University of Queensland

Ian H. DeLacy
The University of Queensland

Peter Wenzl
CIMMYT

Susanne Dreisigacker
CIMMYT

Jose Crossa
CIMMYT

Mark J. Dieters
The University of Queensland

Kaye E. Basford
The University of Queensland and ACPFG

Abstract

Many ecological studies compare the genetic structure of (natural or artificial) populations across different (spatial and/or temporal) environments. Genotyping, using modern high-throughput molecular maker technologies, provides abundant information on (large numbers of) markers and (large numbers of) individuals in populations. An ordering of the markers, based on the relationships among them, enhances the description of the genetic structure of each population. Comparisons of the genetic structures of populations across environments will provide information on the adaptation of the populations. A method is presented for ordering bi-allelic markers for populations of self-fertilising plant species which consist of mixtures of related homozygous genotypes. This method provides stable pair-wise marker similarity measures even when marker frequencies are low, identification of marker combinations that reflect phenomena that cause differentiation (such as selection and migration), and genetic information on the adaptation of the populations to the environments. The method is illustrated using data from a plant breeding program under artificial selection. Here inferences can be made about accumulation of desirable genes (such as for disease resistance), changes in selection objectives, and changes in germplasm in successive populations.

Keywords: marker mapping, plant breeding population, haplotype disequilibrium.
1. Introduction

Plant populations, both natural (undergoing natural selection) and artificial (undergoing artificial selection), are genetically structured in space and time (Beavis 1998; Allard 1999; Stodart et al. 2007) and this structure may be manifested among locations, within locations and among individuals. This genetic structure is the joint result of mutation, migration, selection, drift and reproductive breeding system which all operate within the historical and biological context of each species (Loveless and Hamrick 1984). Thus, the genetic structure of a population reflects its history across time and space and comparing population histories should reveal information on the evolution of adaptedness (Allard 1999). A measure of gametic phase disequilibrium (GPD) can be used to study population history since GPD is affected by any factor causing genetic structure in a population (Falconer and McKay 1996). The assessment of GPD in populations is greatly enhanced by high-throughput marker technology (Fan et al. 2006; Jaccoud et al. 2001) as it enables the investigation of whole-genome GPD.

While not essential for the calculation or study of pairwise GPD among markers, a suitable marker order maximizes the information that can be gleaned from such an investigation. This is particularly so with the use of an increasingly large number of markers (van Os et al. 2005). A marker order is usually obtained from linkage or physical maps. These maps are constructed to reflect GPD due to linkage (i.e. recombination or physical chromosome distance), typically using special (bi-parental) mapping populations. In the study of human populations, a combination of physical maps and GPD measurements has been used to provide information on recombination, selection, population history and gene expression (Tapper et al. 2005). However, for plant studies, existing linkage maps are often unsuitable as each new population requires a different map (Collard et al. 2005) due to differences in the set of polymorphic markers, recombination fractions (Beavis and Grant 1991; Williams et al. 1995) and map resolutions among the populations (Paterson et al. 1990). In addition, physical maps are only available for a limited number of plant species.

Even when suitable linkage or physical maps are available, a requirement exists for a usable marker order that is produced directly from the population under study. This order will reflect all causes of GPD in that population and can be compared across time and space with marker order derived for other populations and with linkage or physical map order when these are available. Many measures of GPD (Hedrick 1987; Devlin and Risch 1995) can be used to order markers. While all of them are based on measures of similarity among markers, most are not stable for low marker frequency (Hubalek 1982) or non-polymorphic markers as the denominator approaches zero. Yet low frequency or non-polymorphic markers are common for populations under selection since desirable genes tend to become fixed over time.

In order to study populations undergoing selection, we propose the use of a similarity measure which is stable under low or non-polymorphic markers and present a method to obtain marker order for bi-allelic marker systems from populations of self-fertilising plant species which consist of mixtures of related homozygous genotypes. The method can be applied to both natural populations (e.g. landraces) and artificial populations (e.g. populations from plant breeding programs) and is illustrated using wheat populations for two generations (i.e. parents and offspring) from an international breeding program.
2. Procedure Development

2.1. Calculating Pairwise GPD

The data produced for a population under study by genotyping a set of \( n_g \) homozygous genotypes with \( n_e \) bi-allelic markers form an \( n_g \times n_e \) binary matrix \( S \) with elements \( s_{ij} \) which record the “state” of the marker as 1 if the genotype has the marker and 0 if it does not. Hence the matrix records whether the genotypes are identical by state, i.e. whether genotypes have the same state for each marker. Any measure of GPD in the population under study is a similarity measure among markers, i.e. the columns in \( S \).

There are many similarity measures for binary data (Hubalek 1982, listed 43) such as that produced by genotyping with bi-allelic markers. These similarity measures are related to each other in that they can be calculated from a two-way contingency table among pairs of markers \( j \) and \( j^t \). Genetically, the contingency table indicates the frequency \( (f_{jj'}) \) of the possible gamete combinations in the population, with the marginal frequencies \( (f_{j\cdot \cdot}, f_{\cdot j'}) \) reflecting expected gamete frequencies under Hardy-Weinberg equilibrium. Some of these similarity measures have been used as a GPD coefficient in genetic studies (e.g. \( r^2 \) and \( D' \); Hedrick 1987; Devlin and Risch 1995). The most commonly used GPD coefficients have the same numerator, the determinant of the \( 2 \times 2 \) contingency table \( (f_{11}f_{22} - f_{12}f_{21}) \), but have different denominators used to standardize the measure (e.g. \( f_{1\cdot \cdot}f_{2\cdot \cdot}/f_{\cdot \cdot \cdot} \) for \( r^2 \) ) (Devlin and Risch 1995). However, all current GPD coefficients measure GPD as a deviation from expected frequencies such as those derived from the application of Hardy-Weinberg equilibrium theory. But these have the problem that they become unstable because their denominator will approach zero when one of the gamete combinations approaches zero. The latter is expected in populations undergoing selection as loci approach fixation.

The Hamann coefficient or the G Index of Agreement (Hamann 1961; Sokal and Sneath 1963; Holley and Guilford 1964) is the difference between the proportions of matches and mismatches in the binary measures on a pair of objects. The Hamann coefficient has been used in psychology, education, taxonomy and social sciences disciplines (Hamann 1961; Sokal and Sneath 1963; Holley and Guilford 1964), but its use in genetics is limited (Leišová et al. 2007). The Hamann coefficient \( g_{jj'} \) between marker \( j \) and \( j^t \) is

\[
g_{jj'} = \left[ (f_{11} + f_{22}) - (f_{12} + f_{21}) \right] / f_{\cdot \cdot \cdot} \tag{1}
\]

where \( f_{11} \) is the frequency of both markers being present. These coefficients form a symmetrical \( n_e \times n_e \) similarity matrix \( G \). Genetically, \( g_{jj'} \) is the difference between coupling and repulsion haplotypes for bi-allelic markers \( j \) and \( j^t \). If the data used to calculate \( g_{jj'} \) were displayed in a \( 2 \times 2 \) contingency table, then the matches are on the diagonal, the mismatches are on the off-diagonals, and the denominator is the sum of the elements in the table. When scored in populations consisting of mixtures of homozygous genotypes, \( g_{jj'} \) is a direct measure of the excess of gamete combinations in coupling \( (f_{11} \text{ and } f_{22}) \) phase (when greater than zero) or repulsion \( (f_{12} \text{ and } f_{21}) \) phase (when less than zero) and reflects all causes of GPD. Thus \( g_{jj'} \) is a measure of observed haplotype disequilibrium (HD), but it does not measure disequilibrium as a deviance from Hardy-Weinberg equilibrium. The \( g_{jj'} \) between two markers \( j \) and \( j^t \) will be one if their patterns over genotypes are identical (coupling) and minus one if they have exactly opposite patterns (repulsion). Importantly, the Hamann coefficient is stable for
Marker Combinations in Natural or Artificial Populations

non-polymorphic markers as its denominator never approaches zero.

Marker phase is important in genetics for the calculation of recombination frequency, as recombination frequency is defined as the number of recombinant gametes over the total number of gamete combinations (Falconer and McKay 1996). Thus, using the absolute value of the Hamann coefficient (\(|g_{jj'}|\)) eliminates the problem of unknown marker phase.

The Hamann coefficient also has a linear relationship with the simple matching coefficient \(m_{jj'}\) between markers \(j\) and \(j'\) (Hubalek 1982):

\[
m_{jj'} = \frac{f_{11} + f_{22}}{f_{11} + f_{22}} = \frac{1}{2}(g_{jj'} + 1).
\] (2)

In bi-parental mapping populations with an expected segregation ratio of 1:1 for each marker (backcross, doubled haploid, \(F_\infty\) populations) the \(m_{jj'}\) has a linear relationship with the recombination fraction \(c_{jj'}\) between markers \(j\) and \(j'\) as

\[
c_{jj'} = 1 - m_{jj'} = \frac{1}{2}(1 - g_{jj'}).
\] (3)

Each \(c_{jj'}\) is a dissimilarity measure and \(m_{jj'}\) is its complementary similarity measure (Gower 1966, 1967). This relationship between \(c_{jj'}\) and \(m_{jj'}\) is not commonly referred to in the molecular mapping literature but was referred to by Hackett (2002). Recognizing that the disequilibrium and recombination measures are complementary similarity and dissimilarity measures means that all standard pattern analysis methodologies, i.e. the joint use of clustering and ordination procedures (Williams 1976; DeLacy et al. 1996), can be directly applied to the appropriate genetic studies.

2.2. Ordering Markers

A two stage procedure was used to order markers: a dendrogram was obtained from a hierarchical clustering, followed by optimization of the marker order along the base of the dendrogram.

In the first stage, a hierarchical agglomerative procedure with \(1 - |g_{jj'}|\) as the dissimilarity measure among markers and group average (or UPGMA, Sokal and Michener 1958) as the clustering strategy was used to form a dendrogram. The use of the absolute value of \(g_{jj'}\) removes the mirror effect due to marker phase and those marker pairs in complete disequilibrium (\(g_{jj'} = 1\) or -1) will be grouped first. Those in equilibrium (no association) will be grouped last. The order of the markers along the dendrogram traces a walk through multidimensional space visiting the position of each marker once only.

However, there are many “marker orders” obtainable from a hierarchical dendrogram as the leaves of the dendrogram can be rotated at any fusion point of the dendrogram. A desirable order would be the shortest possible walk through the disequilibrium space. This is the “travelling salesman problem” and solutions to the problem employ seriation methods (Arabie and Hubert 1996; Hahsler et al. 2008). The seriation method of Gruvaeus and Wainer (1972) as implemented in the seriation package (Hahsler et al. 2008) of the R statistical software (R Development Core Team 2008) was used to optimize marker order. This method uses dendrogram order as scaffolding and flips each leaf of the dendrogram moving up the clustering
so that adjacent entities are the most similar. This algorithm explicitly solves the problem of starting position (i.e. in marker mapping, determining which marker is at the end of the chromosome) by determining which entities are at the extremities. We refer to dendrograms optimized in this manner as “optimized dendrograms”.

This procedure produces an optimized marker order over the whole genome that reflects all processes that cause GPD in the population: selection, migration (founder effect), mutation, drift and linkage. We refer to this order as “genome haplotype disequilibrium order”, or “genome HDO”, to distinguish it from the marker order obtained from a linkage or physical map. The development of this method was motivated by results from microarray studies (Bar-Joseph et al. 2001; Eisen et al. 1998) that have shown genes with similar function tend to group together regardless of their position in chromosomes.

However, markers can also be ordered within each chromosome, referred to as “chromosome HDO”, because it describes GPD within chromosomes. A list of anchor markers allocated to chromosomes can normally be obtained from published comprehensive marker maps (Dodds et al. 2004). If an anchor map which contains sufficient markers used in the study is available, the hierarchical clustering procedure described above can also be used to assign markers with unknown allocation or markers with multiple allocations to the same chromosome as that of the anchor marker(s) with which they first cluster in the dendrogram.

The procedure described here to obtain marker HDO is a modification of the well-known standard mapping procedure that has been implemented in many software packages (e.g. Lander et al. 1987; Stam 1993) to produce genetic linkage maps in plants using bi-parental populations. There are four steps in this standard procedure: (1) calculate the recombination matrix from the observed genotypic data, (2) allocate markers to linkage groups and then to chromosomes, (3) order markers within linkage groups or chromosomes, and (4) calculate map distances. Here we calculate the GPD matrix, use hierarchical clustering to form a dendrogram, order markers across the whole genome, allocate unmapped markers to chromosomes via an anchor map and the dendrogram, and order markers within chromosomes.

If the method is applied to bi-parental mapping populations with expected frequencies of 1:1 for all markers, the chromosome HDO is the same as that derived from standard mapping procedures. In addition, because of the direct relationship between the Hamann coefficient, the simple matching coefficient and the recombination fraction, $1 - |g_{jj'}|$ is a measure of linkage distance along the chromosome and can be converted to centimorgans by the application of the Haldane or Kosambi mapping functions. Hence, the method produces a standard linkage map when applied to standard mapping populations. It has a further advantage in that the absolute value of the Hamann coefficient adjusts for phase differences, so knowledge of the marker status of the parents is not required.

### 2.3. Marker Blocks

A group of adjacent markers that show a high level of disequilibrium is defined as a marker block. The combined use of linkage and/or physical maps and detailed targeted disequilibrium studies enables groups of adjacent markers to be assigned to a marker block if their GPD coefficient exceeds a pre-determined threshold value. Linkage disequilibrium (LD) blocks have been found useful for fine mapping and for identifying recombination hotspots in chromosome
regions (Maniatis et al. 2002; Tapper et al. 2005). Similarly, the absolute value of the Hamann coefficient, \(|g_{jj}'|\), though limited to dominant bi-allelic marker systems, can be employed, in conjunction with maps, to identify marker blocks.

A haplotype disequilibrium block (HDB) is defined as a group of adjacent markers (or a single marker) with the absolute value of the Hamann coefficient greater than or equal to a threshold. The threshold value indicates the minimum excess of the most common pair of haplotypes over the other types and the value chosen will depend on the structure of the populations. Using a higher threshold increases the likelihood that the identified HDBs are due to linkage. However, in addition to linkage, HDBs indicate all the processes (selection, founder (migration), mutation, genetic drift) that affect GPD in the population under study. They will be common in artificial or natural populations that are highly selected.

3. Case Study

The procedure described above was applied to two successive populations of inbred wheat lines from the Elite Spring Wheat Yield Trial (ESWYT) at the International Maize and Wheat Improvement Center (CIMMYT) in México. The first population consisted of 685 entry lines tested in the first 25 cycles of the ESWYT (ESWYT entries) and the second population consisted of 195 parental lines (ESWYT parents), with an average of six years gap between the two successive populations. Not all parents of the entries were included in that second population. Most of these lines were bred at CIMMYT for low latitude, irrigated, high input conditions. These lines form an artificial population of genealogically connected small families of inbred lines.

Of the 685 ESWYT entries, 599 (plus some duplicates) were genotyped and a total of 1,447 polymorphic DArT markers were scored (data provided by CIMMYT). The 195 parents were genotyped and 2,153 polymorphic DArT markers were scored (data provided by CIMMYT). The reasons for observing different markers were that there were different levels of polymorphism between the parents and the entries and non-polymorphic markers were not recorded. Both chromosome HDO and genome HDO were constructed for each of the populations. There were 741 markers in common, enabling a comparison of HDO for these successive populations of this breeding program. This enables a description of the change in the genetic structure of the ESWYT germplasm over time. A threshold value of 0.8 for the absolute value of the Hamann coefficient, \(|g_{jj}'|\), was used to define a HDB in both populations.

An anchor map was derived from the Synthetic×Opata double haploid mapping population (Sorrells et al. 2011). This is one of the most dense DArT maps published for wheat and it provided 545 and 457 anchor markers for the ESWYT entries and the ESWYT parent populations, respectively.

4. Results

The map for the Synthetic×Opata double haploid mapping population obtained using the method described here was essentially the same as that produced by the standard mapping software, EasyMap (not shown here). It confirmed that using the absolute value of the
Hamann coefficient successfully deals with phase differences for marker mapping.

An additional 857 markers for the entries and 1145 markers for the parents were successfully allocated to chromosomes. This enabled $545 + 857 = 1402$ markers to be ordered for the entries under chromosome HDO and $457 + 1145 = 2602$ markers to be ordered for the parents under chromosome HDO. Eighty-four (11%) of the 741 markers common to both generations were assigned to different chromosomes. In addition, 31 markers (25 of them common with markers for the parents) that had previously been allocated to chromosome 1B were reassigned to the 1BL/1RS translocation (data not shown) based on the haplotype profiles of 32 released cultivars included in the ESWYT population with known 1BL/1RS translocation status (Zeller 1973).

Chromosome HDO reflects HD within each chromosome, whereas genome HDO reflects HD across the whole genome (Figure 1). Long-distance HD was detected in both the parents (Figure 1a) and the entries (Figure 1c) and was not caused by linkage (as it was across different chromosomes). In such a plant breeding population, this HD is likely to be the result of selection. As the markers in the 1BL/1RS translocation were in disequilibrium with markers in chromosome 1BS, they were ordered together in genome HDO. In the ESWYT population, $r^2$ and $D'$ detected less long-distance disequilibrium than the Hamann coefficient (Figure 2).

Two types of HDB, chromosome HDB based on chromosome HDO (Figures 1a and 1c) and genome HDB based on genome HDO (Figures 1b and 1d) were identified. There were more chromosome HDBs (631 and 621) than genome HDBs (371 and 380) for both the parents and the entries. The smaller number of genome HDBs was due to the merging of several chromosome HDBs into a single genome HDB. The biggest genome HDB in the parents (Figure 1b red triangle) consisted of 17 chromosome HDBs, while the biggest one in the entries (Figure 1d) consisted of 19 chromosome HDBs. The biggest genome HDBs in both parents and entries were not identified using $r^2$ and $D'$ (Figures 2b and 2d). Low or non-polymorphic markers in these blocks were not adequately represented.

Chromosome and genome HDOs constructed from the parents were different from those obtained from the entries (Figure 3), both in order and in size. While most chromosome HDBs were retained in both generations, the relative orientation of these blocks was different (Figure 3). In several cases, chromosome HDBs in the parents were merged into a single chromosome HDB in the entries, but the reverse was rare (Figure 3). For example, the two HDBs in chromosome 2A of the parents were merged into a single HDB in the entries (Figure 3).
Figure 1: Graphical depiction of the matrices of the absolute value of the Hamann coefficient ($|g_{jj'}|$) among markers for two generations of the Elite Spring Wheat Yield Trial (ESWYT) population based on chromosome and genome haplotype disequilibrium order (HDO). Graphical depiction of $|g_{jj'}|$ for the parents based on (a) chromosome and (b) genome HDO and for the entries based on (c) chromosome and (d) genome HDO. Horizontal lines in the chromosome HDO indicate chromosomes. Arrows indicate the re-arrangement of markers between chromosome and genome HDO for markers in the 1BL/1RS translocation and the short arm of chromosome 1B.
Figure 2: Graphical depiction of the matrices of the $|D'|$ and $r^2$ coefficients among markers for two generations of the Elite Spring Wheat Yield Trial (ESWYT) population based on chromosome and genome haplotype disequilibrium order (HDO). Graphical depiction of $|D'|$ and $r^2$ for the parents based on (a) chromosome and (b) genome HDO and for the entries based on (c) chromosome and (d) genome HDO. Horizontal lines in the chromosome HDO indicate chromosomes. Arrows indicate markers in the 1BL/1RS translocation and the short arm of chromosome 1B. Unlabeled black square in genome HDO indicates the biggest genome HD block (HDB) identified using the Hamann coefficient.
Figure 3: Chromosome haplotype disequilibrium order (HDO) constructed using two generations of the Elite Spring Wheat Yield Trial (ESWYT) population. Comparison of chromosome HDO from the parents (black) and the entries (purple) for (a) all markers and (b) markers in chromosome 2A for the parents (left) and the entries (right).
5. Discussion

The procedure described in this paper to obtain marker HDO involved (1) calculating pairwise GPD using the Hamann coefficient, (2) using hierarchical clustering to form a dendrogram, (3) ordering markers across the whole genome, (4) allocating unmapped markers to chromosomes via an anchor map and the dendrogram, and (5) ordering markers within chromosomes. It can be applied to any population consisting of related homozygous genotypes, to produce marker HDOs based on the similarity of bi-allelic marker combinations across genotypes. The haplotypes among members of a population reflect disequilibrium, as a set of markers will have identical patterns across genotypes if they are either co-located (linked) or the only marker combinations in the founders, are selected together, or a combination of these.

Two types of marker order are defined: chromosome and genome haplotype disequilibrium order (HDO). These two types of HDO provide complementary information on population history. Chromosome HDO provides information on chromosome haplotype disequilibrium blocks (HDBs) that are more influenced by linkage, while genome HDO is used to identify genome haplotype disequilibrium blocks (HDBs) that are affected by any factor causing GPD. For example, the HDB in the 1BL/1RS translocation (Figure 1) is definitely caused by linkage since there is no recombination in this translocation (Lukaszewski 2000).

Markers with low or no haplotype variability are expected in highly selected populations, such as the ESWYT. These markers will group together in either chromosome or genome HDBs. While many HDBs were identified using the absolute value of the Hamann coefficient, most were not identified using the two common GPD coefficients, $D'$ and $r^2$ (Figure 2). These two common GPD coefficients identified a marker block in the 1BL/1RS translocation that was carried by half of the ESWYT lines, but did not identify the biggest genome HDB in the ESWYT that had very low haplotype variability.

Comparison of HDO from populations that differ in space and time enables a study of the evolution of adaptedness. This approach will be more straightforward and useful if both populations have the same set of markers scored, whether or not they are polymorphic. This is because low or non-polymorphic markers in such studies provide information on marker fixation due to selection during evolution. Many smaller HDBs in the parents were amalgamated in the entries (Figure 3) implying that such blocks were the result of selection over generations.

HDO produced from special bi-parental populations with no selection and expected Mendelian segregation ratio of 1:1 (e.g. double haploid, backcross, $F_{\infty}$) reflects only GPD due to linkage. When applied to such a population, the method outlined here produces essentially the same map order as standard mapping procedures and does not require parental information. In addition, the relationships among Hamann, simple matching coefficient and recombination frequency can be used to calculate map distance and produce a standard linkage map. Moreover, the threshold to determine HDB in such special populations can be adjusted. For example, in the Synthetic×Opata double haploid population a threshold of $|g_{ij}| \geq 0.6$ identified genome HDBs that corresponded to a chromosome arm (data not shown).

The procedure described here enables the determination of a marker order that is useful for any population that consists of inter-related inbred lines. HDBs can be identified to study
the population structure. Comparing HDO with linkage or physical order provides additional information about the cause of GPD in the population. This procedure can be used to study the evolution of any artificial or natural population as an appropriate HDO can be generated for any temporal and/or spatial partition of a dynamic population.

Acknowledgements

The first author thanks the Endeavour International Postgraduate Research Scholarship for the financial support during the course of her PhD Studies. Financial support from The University of Queensland and CIMMYT is also gratefully acknowledged. We wish to thank Mark Sorrells at Cornell University for supplying the Synthetic×Opata map.

References


Marker Combinations in Natural or Artificial Populations


Affiliation:

Vivi N. Arief
The University of Queensland, School of Agriculture and Food Sciences
Brisbane, QLD 4072
E-mail: v.arief1@uq.edu.au

Ian H. Delacy
The University of Queensland, School of Agriculture and Food Sciences
Brisbane, QLD 4072
E-mail: i.delacy@uq.edu.au

Peter Wenzl
International Maize and Wheat Improvement Center (CIMMYT)
Apdo. Postal 6-641, 06600, Mexico, DF, Mexico
E-mail: p.wenzl@cgiar.org

Susanne Dreisigacker
International Maize and Wheat Improvement Center (CIMMYT)
Apdo. Postal 6-641, 06600, Mexico, DF, Mexico
E-mail: s.dreisigacker@cgiar.org

Jose Crossa
International Maize and Wheat Improvement Center (CIMMYT)
Apdo. Postal 6-641, 06600, Mexico, DF, Mexico
E-mail: j.crossa@cgiar.org

Mark J. Dieters
The University of Queensland, School of Agriculture and Food Sciences
Brisbane, QLD 4072
E-mail: m.dieters@uq.edu.au

Kaye E. Basford
The University of Queensland, School of Agriculture and Food Sciences and
Australian Centre for Plant Functional Genomics (ACPFG)
Brisbane, QLD 4072
E-mail: k.basford@uq.edu.au