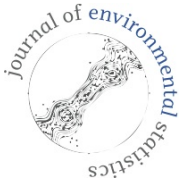

Journal of Environmental Statistics

February 2013, Volume 4, Issue 4. http://www.jenvstat.org

Using Molecular Marker Order to compare Genetic Structure in Plant Populations undergoing Selection

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 populations undergoing natural or artificial selection across different environments. High-throughput molecular markers are now commonly used for these comparisons and provide information on the adaptation of the populations to their environments. The genetic structure reflects the history of selection, mutation, migration, and the reproductive breeding system of the populations in their environments. This can be investigated by comparing the ordering of markers obtained from the population with that provided by a recombination or physical map. In populations undergoing selection many genes (markers) have low or zero frequency and commonly used disequilibrium coefficients become unstable under these conditions. A method is presented for ordering bi-allelic markers for populations of self-fertilizing plant species which consist of mixtures of related homozygous genotypes. This 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 and inferences are made about accumulation of desirable genes (such as for disease resistance).

Keywords: plant breeding and natural populations, haplotype disequilibrium.

1. Introduction

Plant populations, both those undergoing natural and 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). For the study of plant populations, marker order that reflects all causes of GPD can be directly constructed for each population using disequilibrium measures. Comparing this marker order across populations and/or with linkage or physical maps enables the study of changes in the populations due to selection.

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 frequency or non-polymorphic markers as the denominator approaches zero (Hubalek 1982). Yet low frequency or non-polymorphic markers are common for populations undergoing 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 geno-

types with n_e bi-allelic markers form an $n_g \times n_e$ matrix $\boldsymbol{S}$ 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 $\boldsymbol{S}$.

There are many similarity measures for binary data (43 were listed by Hubalek 1982) such as those 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'. Genetically, the contingency table indicates the frequency ($f_{jj\prime}$) of the possible gamete combinations in the population, with the marginal frequencies ($f_{j\bullet}, f_{\bullet j\prime}$) 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 2×2 contingency table ($f_{11}f_{22} - f_{12}f_{21}$), but have different denominators used to standardize the measure (e.g. $f_{1\bullet}f_{2\bullet}f_{\bullet 1}f_{\bullet 2}$ 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 $g_{jj\prime}$ between marker j and j' is

$$g_{jj\prime} = [(f_{11} + f_{22}) - (f_{12} + f_{21})]/f_{\bullet\bullet} \tag{1}$$

where f_{11} is the frequency of both markers being present. These coefficients form a symmetrical $n_e \times n_e$ similarity matrix $\boldsymbol{G}$. Genetically, $g_{jj\prime}$ is the difference between coupling and repulsion haplotypes for bi-allelic markers j and j'. If the data used to calculate $g_{jj\prime}$ were displayed in a 2×2 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, $g_{jj\prime}$ is a direct measure of the excess of gamete combinations in coupling (f_{11} and f_{22}) phase (when greater than zero) or repulsion (f_{12} and f_{21}) phase (when less than zero) and reflects all causes of GPD. Thus $g_{jj\prime}$ is a measure of observed haplotype disequilibrium (HD), but it does not measure disequilibrium as a deviance from Hardy-Weinberg equilibrium. The $g_{jj\prime}$ between two markers j and j' 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 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 ($|g_{jj\prime}|$) eliminates the problem of unknown marker phase.

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

$$m_{jj\prime} = (f_{11} + f_{22})/f_{\bullet\bullet} = (g_{jj\prime} + 1)/2 \ . \tag{2}$$

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

$$c_{jj\prime} = 1 - m_{jj\prime} = (1 - g_{jj\prime})/2 \ . \tag{3}$$

Each $c_{jj\prime}$ is a dissimilarity measure and $m_{jj\prime}$ is its complementary similarity measure (Gower 1966, 1967). This relationship between $c_{jj\prime}$ and $m_{jj\prime}$ 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 $1 - |g_{jj\prime}|$ 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 $g_{jj\prime}$ removes the mirror effect due to marker phase and those marker pairs in complete disequilibrium ($g_{jj\prime} = \pm 1$) 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 "traveling 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 Core Team 2012) 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 containing 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, migration, mutation, and genetic drift) that affect GPD in the population under study. They will be common in artificial or natural populations that are undergoing selection.

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). It is estimated that there is an average six-year gap between the two generations as six years is the time required in the CIMMYT what breeding cycle to develop the best advance lines after making new crosses (Wang *et al.* 2003). The entries and parents were mostly bred at CIMMYT and targeted to low latitude, irrigated, high input conditions. These lines form a population of genealogically connected small families of inbred lines.

Five hundred and ninety-nine (plus some duplicates) of the 685 ESWYT entries were genotyped using composite DArT chip v2.3 and 1,447 polymorphic DArT markers were scored. About two years later, the 195 parents were genotyped using DArT chip v2.6 and 2,153 polymorphic DArT markers were scored. DArT chip v2.6 contains 5,000 markers including the 2,500 markers from DArT chip v2.3 (`http://www.trticarte.com.au/content/wheat_diversity_analysis.html`), hence more markers were expected to be scored on the parental population. As low or non-polymorphic markers in both populations were not scored, only 741 polymorphic markers common to both populations were available for comparing the HDO across the two generations of this breeding program. Both chromosome HDO and genome HDO were constructed for each of the populations and used to describe the change in genetic structure of the ESWYT population over time. Following the procedure of defining LD blocks in human studies (Maniatis *et al.* 2002), a threshold value of 0.8 for the absolute value of the Hamann coefficient, $|g_{jj\prime}|$, was used to define a HDB in both populations. The wheat consensus map (Huang *et al.* 2012) with 4,606 DArT markers was used as an anchor map and provided 1,115 and 1,596 anchor markers for the ESWYT entries and parents, respectively. The HDO from both ESWYT populations were compared to the order of this consensus map.

The procedure described above was also applied to a bi-parental mapping population, the Synthetic×Opata (SO) double haploid mapping population (Sorrells *et al.* 2011). This population consisted of 163 double haploid lines with a total of 1,414 polymorphic DArT markers. The chromosome HDO was then compared to the published SO map produced using a standard mapping software EasyMap (Sorrells *et al.* 2011) .

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 299 markers for the entries and 469 markers for the parents were successfully allocated to chromosomes. These included 31 markers (25 of them common with markers for the parents) that were reassigned to the 1BL/1RS translocation based on the haplotype profiles of 32 released cultivars included in the entries with known 1BL/1RS translocation status (data not shown). 1BL/1RS is a substitution of a rye chromosome arm for the short arm of chromosome 1B (Zeller 1973). This translocation, observed on about half of the entries, provides several rust resistance genes and has contributed to high-yielding genotypes. In total, there were 1,414 and 2,056 markers ordered for the entries and parents, respectively.

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 a plant breeding population, this long-distance HD is likely to be the result of selection. As the markers in the 1BL/1RS translocation were in disequilibrium with markers in the short arm of chromosome 1B, they were ordered together in genome HDO. In the SO population, the three disequilibrium coefficients produced similar results and detected no long-distance HD as this population was not undergoing selection (data not shown). In the ESWYT parental and entries populations, r^2 and D' detected less long-distance HD than

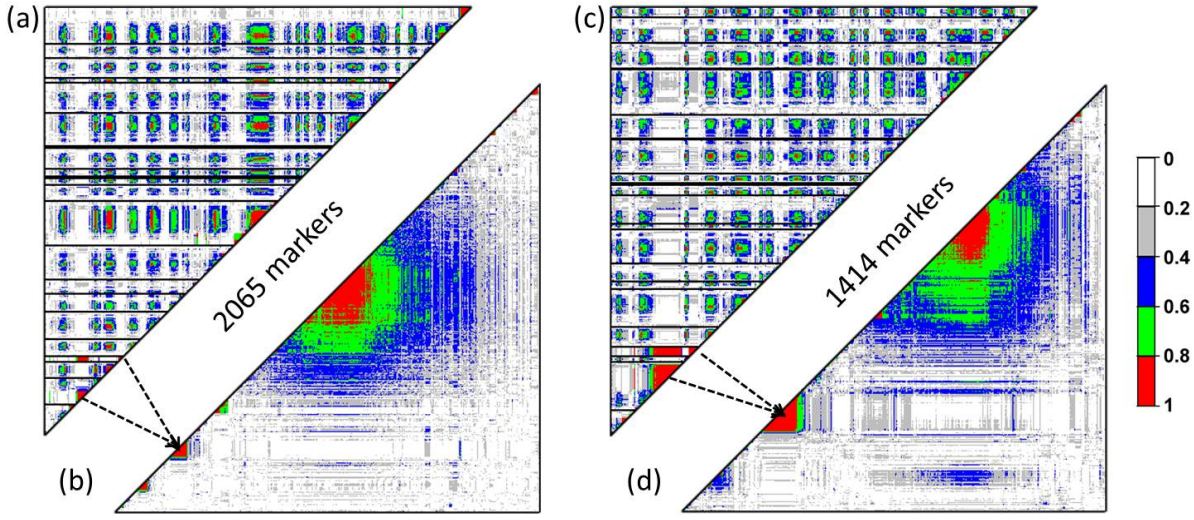

Figure 1: Graphical depiction of the matrices of the absolute value of the Hamann coefficient ($|g_{jj\prime}|$) among markers for two generations of the Elite Spring Wheat Yield Trial (ESWYT) population based on chromosome and genome haplotype disequilibrium order (HDO). Graphical depiction of $|g_{jj\prime}|$ 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.

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 (727 and 485) than genome HDBs (691 and 448) for both the parents

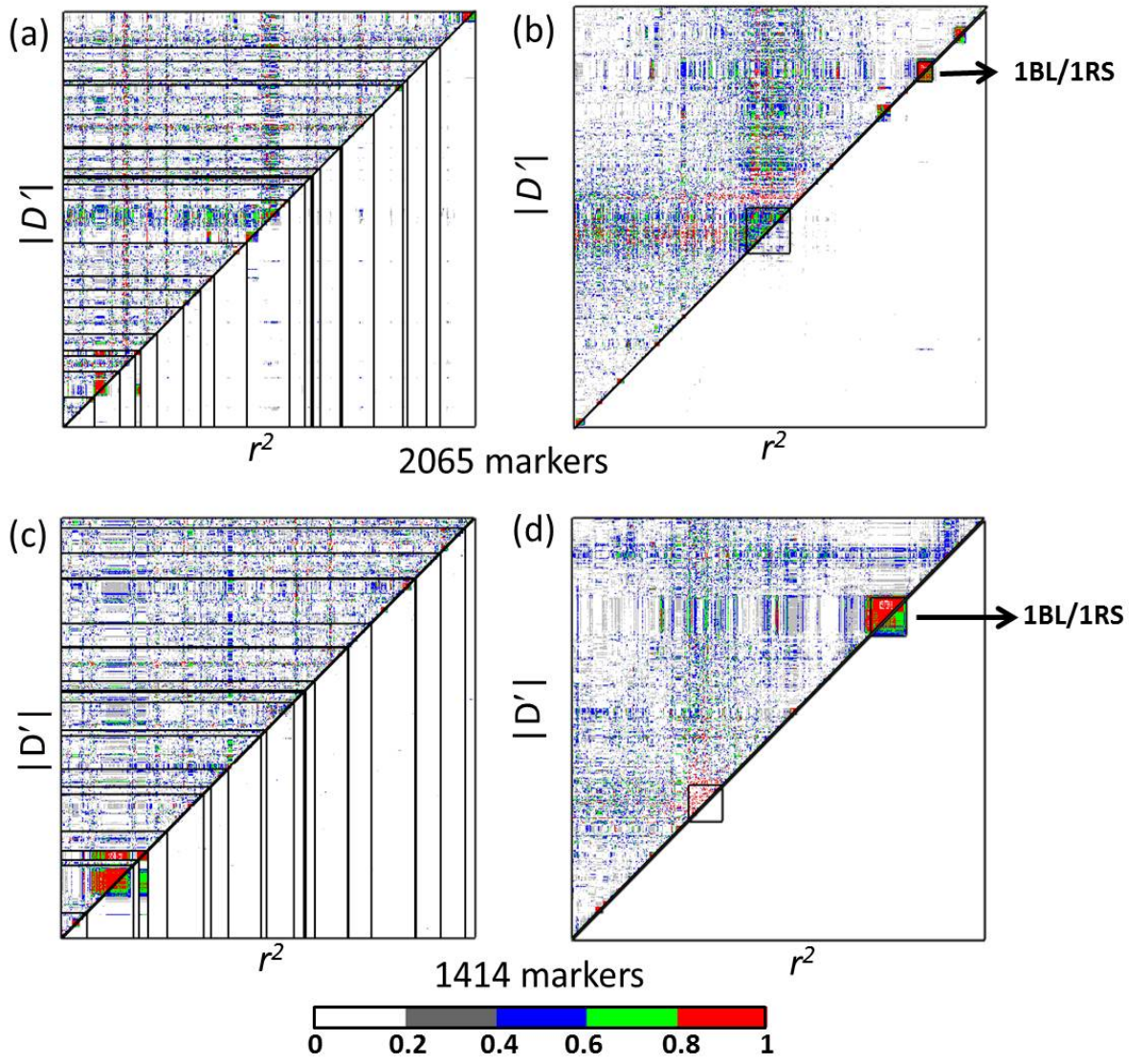

Figure 2: Graphical depiction of the matrices of the $|D'|$ and r^2 coefficients among markers for two generations of the Elite Spring Wheat Yield Trial (ESWYT) population based on chromosome and genome haplotype disequilibrium order (HDO). Graphical depiction of $|D'|$ and r^2 for the parents based 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.

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 18 chromosome HDBs, while the biggest one in the entries (Figure 1d) consisted of 17 chromosome HDBs. The biggest genome HDBs in both parents and entries were not identified using r^2 and D' (Figures 2b and 2d) as these blocks consisted of low polymorphic markers. In the SO population, both chromosome and genome HDBs reflect the chromosomes arms and were identified using the three GPD coefficients (data not shown).

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

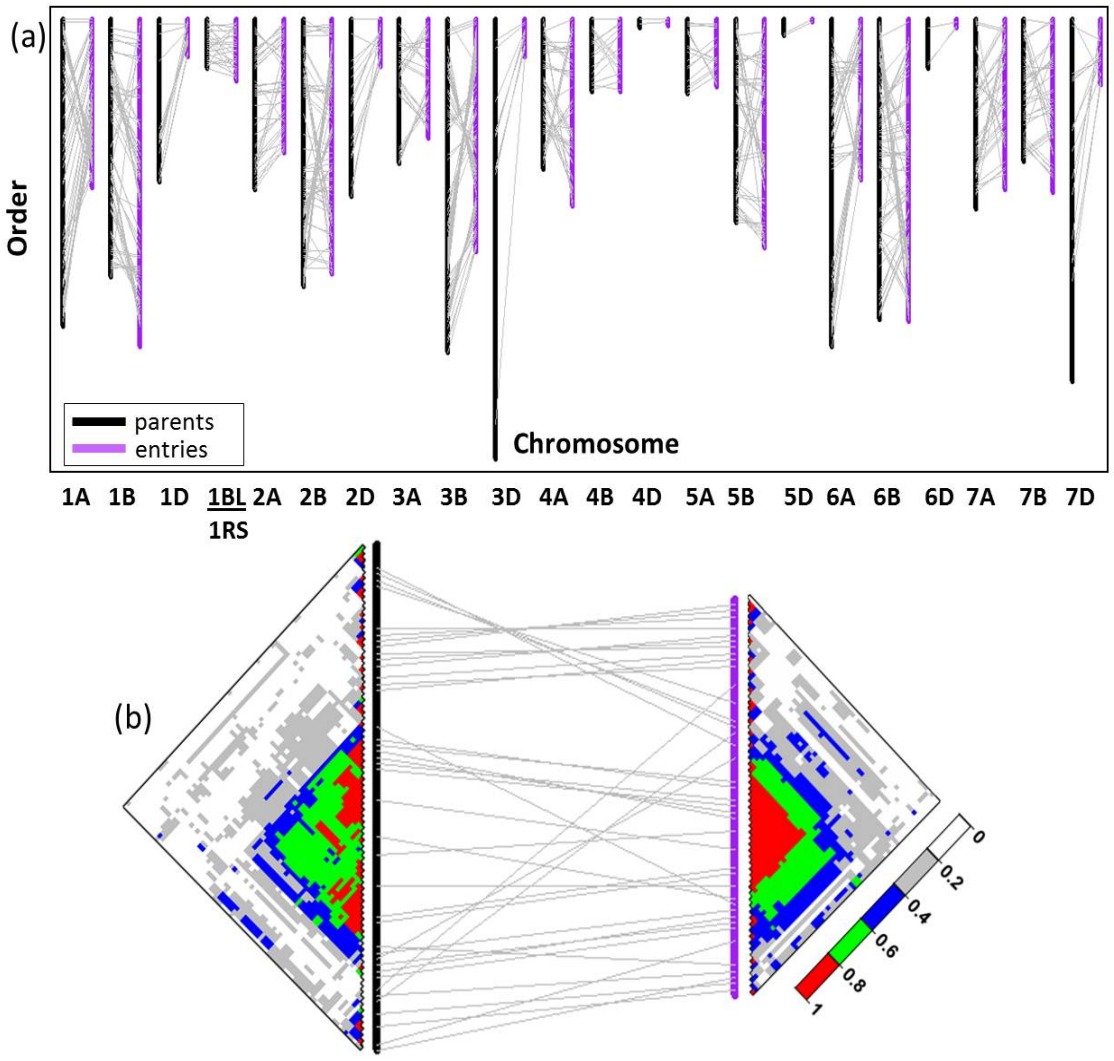

Figure 3: Chromosome haplotype disequilibrium order (HDO) constructed using two generations of the Elite Spring Wheat Yield Trial (ESWYT) population. Comparison of chromosome HDO from the parents (black) and the entries (purple) for (a) all markers and (b) markers in chromosome 2A for the parents (left) and the entries (right).

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 chromo-

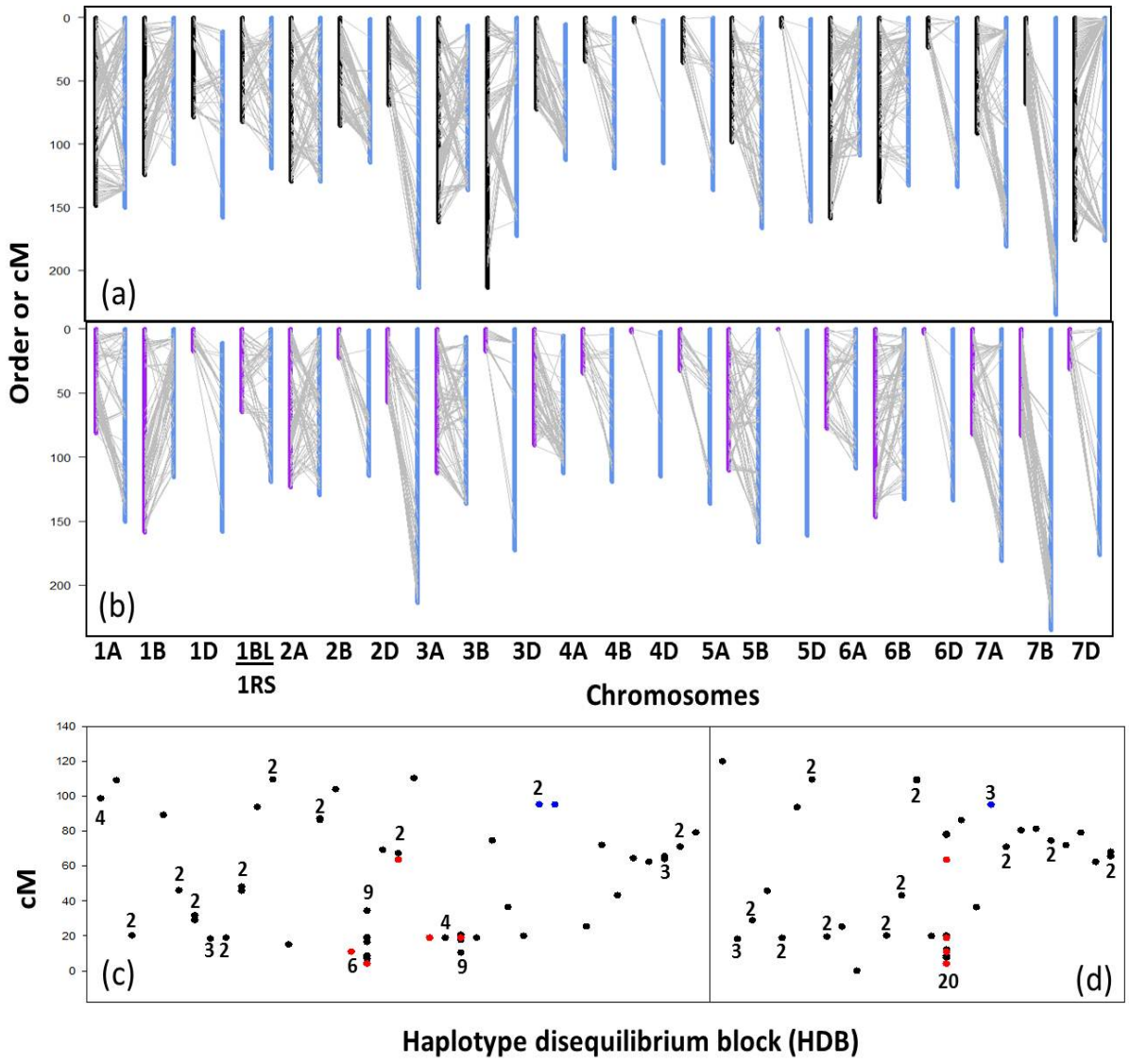

Figure 4: Chromosome haplotype disequilibrium order (HDO) constructed using two generations of the Elite Spring Wheat Yield Trial (ESWYT) and DArT consensus map (Huang *et al.* 2012). (a) Comparison of chromosome HDO from the parents (blacks) and the consensus map (blue). (b) Comparison of chromosome HDO from the entries (purple) and the consensus map (blue). Markers in the consensus map were displayed in cM, while markers in both ESWYT populations were displayed as order. Consensus map distance (cM, vertical axis) for chromosome haplotype disequilibrium blocks (HDBs) in chromosome 2A arranged in chromosome HDO (horizontal axis) from the parents (c) and the entries (d). The number of markers in each HDB was displayed, except for single-marker HDBs. The colored points indicated the markers that were common to the parental and entry populations and discussed in the text.

some 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). For genome HDO, the biggest HDB in the entries consisted of 20 smaller HDBs in the parents with some adjacent blocks. The re-ordering and re-sizing of HDBs between the parents and the entries were likely to reflect selection that occurred over time between the parental and entry populations. For example, the biggest genome HDB in the entries included markers from chromosome 2A and 7D. These chromosomes were known to have, respectively, the photoperiod gene *Ppd3* (Worland *et al.* 1998) that is important for adaptation and the *Lr19* translocation that carries a leaf rust resistance gene (Sharma and Knott 1966). In the parents, markers for each chromosome were allocated to HDBs which were not adjacent to one another. This result could indicate the effect of selection for the *Ppd3* and *Lr19* genes.

With few exceptions, the markers in a chromosome HDB were co-located or close together in the consensus map (Figure 4). For example, a HDB in the entry population (Figure 4d, red points) consisted of two clusters of markers at 4-20cM and 63-78cM. This HDB is likely to be a result of selection. In the parental population, the markers in this HDB formed five smaller HDBs (Figure 4c, red points), but with selection they coalesced. Some co-located or closely linked (<10cM) markers in the consensus map were divided into several adjacent HDBs in the parental population (Figure 4c, blue points). This might be an indication that the threshold value used in defining the HDBs in the parents was too high, especially since these markers grouped into a single HDB in the entries (Figure 4d, blue points).

5. Discussion

The procedure described here, using the Hamann coefficient ($|g_{jj'}|$), enables the determination of a marker order that is useful for studying the disequilibrium of any population that consists of inter-related inbred lines, especially those under selection. The standard coefficients, D' and r^2, have a limitation when studying populations under selection as they become unstable when the frequency of alleles approaches zero (the denominator approaches zero) and they are undefined at zero frequencies. Low or non-polymorphic allele frequencies must be expected when comparing populations under selection. Comparing HDO with linkage order from a consensus or physical map provides information about the cause of GPD in the population. In addition, HDBs can be identified and used to study the population structure. As an appropriate HDO can be generated for any temporal and/or spatial partition of an artificial or natural population, this procedure facilitates study of evolution.

While the development of the Hamann coefficient given here is restricted to bi-allelic loci, this coefficient can be generalized to multi-allelic systems by averaging all the pairwise coefficients among alleles as is done with D' and r^2.

The procedure 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, or they are selected together, or a combination of these.

The derived chromosome and genome HDO provide complementary information on population history (Figure 2). Chromosome HDO enables the identification of chromosome HDBs that are likely to be influenced by linkage, while genome HDO is used to identify genome HDBs that are likely to be affected by any factor causing GPD. For example, the HDB consisting of markers on the BL/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 commonly used 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 polymorphic markers. A pair of markers with low polymorphism will have a small value of D' and an even smaller value of r^2 and these coefficients will not be defined if there is no polymorphism. While it is common practice to remove low or non-polymorphic markers from the analysis, for populations undergoing selection, these low or non-polymorphic markers indicate regions of chromosomes that are highly selected and are reaching or have reached fixation. For plant breeding purposes, the information on low or non-polymorphic markers in the breeding populations can be useful in designing the crossing strategies, evaluating the effectiveness of selection, and measuring the genetic diversity in the breeding populations.

The recommended measure of disequilibrium enables a comparison of HDO from populations that differ in space and time, enhancing studies of the evolution of adaptedness as outlined by Allard (1999). This approach will be most useful if all populations have the same set of markers scored, whether or not they have low or no polymorphism as these provide information on marker fixation due to selection during evolution.

HDO produced from special bi-parental populations with no selection and expected Mendelian segregation ratio of 1:1 (e.g. double haploid, backcross, and F_∞) 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 this case, the relationships among the Hamann coefficient, 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, e.g. in the Synthetic×Opata double haploid population we used a threshold value of 0.6 and 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 consisting of inter-related inbred lines. Thus, it can be used to study changes in natural or artificial populations undergoing selection even when low or non-polymorphic markers are present.

Acknowledgments

The first author thanks the Endeavor International Postgraduate Research Scholarship for 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 data.

References

Allard RW (1999). *Principles of Plant Breeding, 2nd Edition*. John Wiley and Sons, New York.

Arabie P, Hubert LJ (1996). *An Overview of Combinatorial Data Analysis*, pp. 5–63. World Scientific, Singapore.

Bar-Joseph Z, Gifford DK, Jaakkola TS (2001). "Fast optimal Leaf Ordering for Hierarchical Clustering." *Bioinformatics*, **17**(Suppl. 1), S22–S29.

Beavis WD (1998). *QTL Analyses: Power, Precision, and Accuracy*, pp. 145–162. CRC Press, Boca raton.

DeLacy IH, Basford KE, Cooper M, Fox PN (1996). *Restropective Analysis of Historical Data Sets from Multi-environmental Trials - Theoretical Development*, pp. 243–267. CAB International, Wallingford, UK.

Devlin B, Risch N (1995). "A Comparison of Linkage Disequilibrium Measures for Fine-scale Mapping." *Genomics*, **29**, 311–322.

Dodds KG, Ball R, Djorovic N, Carson SD (2004). "The Effect of an Imprecise Map on Interval Mapping QTLs." *Genetical Research*, **84**, 47–55.

Eisen MB, Spellman PT, Brown PO, Botstein D (1998). "Cluster Analysis and Display of Genome-wide Expression Patterns." *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 14863–14868.

Falconer DS, McKay TFC (1996). *Introduction to quantitative genetics*. 4th edition. Longman, Burnt Mill, England.

Fan JB, Chee MS, Gunderson KL (2006). "Highly Parallel Genomic Assays." *Nature Reviews Genetics*, **7**, 632–644.

Gower JC (1966). "Some Distance Properties of Latent Root and Vector Methods Used in Mutivariate Analysis." *Biometrika*, **53**(3/4), 325–338.

Gower JC (1967). "Multivariate Analysis and Multidimensional Geometry." *Statistician*, **17**(1), 13–28.

Gruvaeus G, Wainer H (1972). "Two Additions to Hierarchical Cluster Analysis." *British Journal of Mathematical and Statistical Psychology*, **25**, 200–206.

Hackett CA (2002). "Statistical Methods for QTL Mapping in Cereals." *Plant Molecular Biology*, **48**, 585–599.

Hahsler M, Hornik K, Buchta C (2008). "Getting Things in Order: An Introduction to the R Package *seriation*." *Journal of Statistical Software*, **25**(3), http://www.jstatsoft.org/v25/i03/paper.

Hamann U (1961). "Merkmalsbestand und Verwandtschaftsbeziehungen der Farinosae: Ein Beitrag zum System der Monokotyledonen." *Willdenowia*, **2**(5), 639–768.

Hedrick PW (1987). "Genetic Disequilibrium Measures: Proceed with Caution." *Genetics*, **117**, 331–341.

Holley J, Guilford J (1964). "A Note on the G Index of Agreement." *Educational and Psychological Measurement*, **24**, 749–753.

Huang BE, George AW, Forrest KL, Kilian A, Hayden MJ, Morell MK, Cavanagh CR (2012). "A multiparent advanced generation inter-cross population for genetic analysis in wheat." *Plant Biotechnology Journal*, **10**, 826–839.

Hubalek Z (1982). "Coefficients of Association and Similarity, Based on Binary (Presence-absence) Data: An Evaluation." *Biological Review*, **57**, 669–689.

Jaccoud D, Peng K, Feinstein D, Kilian A (2001). "Diversity Arrays: A Solid State Technology for Sequence Information Independent Genotyping." *Nucleic Acids Research*, **29**(4), e25.

Lander ES, Green P, Abrahamson J, Barlow A, Daly MJ, Lincoln SE, Newburg L (1987). "MAPMAKER: An Interactive Computer Package for Constructing Primary Genetic Linkage Maps of Experimental and Natural Populations." *Genomics*, **1**, 174–181.

Leišová L, Kučera L, Dotlačil L (2007). "Genetic Resources of Barley and Oat Characterised by Microsatellites." *Czech Journal of Genetic and Plant Breeding*, **43**(3), 97–104.

Loveless MD, Hamrick JL (1984). "Ecological Determinants of Genetic Structure in Plant Populations." *Annual Review of Ecology and Systematics*, **15**, 65–95.

Lukaszewski AJ (2000). "Manipulation of the 1RS.1BL Translocation in Wheat by Induced Homoeologous Recombination." *Crop Science*, **40**, 216–225.

Maniatis N, Collins A, Xu CF, McCarthy LC, Hewett DR, Tapper W, Ennis S, Ke X, Morton NE (2002). "The First Linkage Disequilibrium (LD) Maps: Delineation of Hot and Cold Blocks by Diplotype Analysis." *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 2228–2233.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Sharma D, Knott D (1966). "The transfer of leaf rust resistance from Agropyron to Triticum by irradiation." *Canadian Journal of Genetics and Cytology*, **8**, 137–143.

Sokal RR, Michener CD (1958). "A Statistical Method for Evaluating Systematic Relationships." *University of Kansas Science Bulletin*, **38**, 1409–1438.

Sokal RR, Sneath PHA (1963). *Principles of Numerical Taxonomy*. W. H. Freeman and Company, San Fransisco.

Sorrells ME, Gustafson JP, Somers D, Chao S, Benscher D, Guedira-Brown G, Huttner E, Kilian A, McGuire PE, Ross K, Tanaka J, Wenzl P, Williams K, Qualset CO (2011). "Reconstruction of the Synthetic W7984 ×Opata M85 Wheat Reference Population." *Genome*, **54**(11), 875–882.

Stam P (1993). "Construction of Integrated Genetic Linkage Maps by Means of a New Computer Package: JoinMap." *Plant Journal*, **3**(5), 739–744.

Stodart BJ, Mackay MC, Raman H (2007). "Assessment of Molecular Diversity in Landraces of Bread Wheat (*Triticum aestivum* L.) held in an Ex Situ Collection with Diversity Arrays Technology (DArT^{TM})." *Australian Journal of Agricultural Research*, **58**, 1174âĂŞ1182.

Tapper W, Collins A, Gibson J, Maniatis N, Ennis S, Morton NE (2005). "A Map of the Human Genome in Linkage Disequilibrium Units." *Proceedings of the National Academy of Sciences of the United States of America*, **102**(33), 11835–11839.

van Os H, Stam P, Visser RGF, van Eck HJ (2005). "SMOOTH: A Statistical Method for Successful Removal of Genotyping Errors from High-density Genetic Linkage Data." *Theoretical and Applied Genetics*, **112**, 187–194.

Wang J, van Ginkel M, Trethowan R, Pfeiffer W (2003). "Documentation of the CIMMYT Wheat Breeding Programs." *Technical report*, Wheat Program CIMMYT.

Williams WT (1976). *Pattern Analysis in Agricultural Science*. Elsevier Scientific Publishing Company, Amsterdam.

Worland AJ, Boerner A, Korzun V, Li WM, Petrovic S, Sayers EJ (1998). "The influence of photoperiod genes on the adaptability of European winter wheats." *Euphytica*, **100**, 395–394.

Zeller FJ (1973). "1B/1R Wheat-Rye Chromosome Substitutions and Translocation." In ER Sears, LMS Sears (eds.), *4th International Wheat Genetics Symposium*, pp. 209–221. University of Misourri.

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, México, DF, México
E-mail: p.wenzl@cgiar.org

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

Jose Crossa
International Maize and Wheat Improvement Center (CIMMYT)
Apdo. Postal 6-641, 06600,México, DF, México
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

Journal of Environmental Statistics
Volume 4, Issue 4
February 2013

http://www.jenvstat.org
Submitted: 2012-04-10
Accepted: 2012-10-05