MARKER ASSISTED BREEDING IN SUGARCANE: A COMPLEX POLYPLOID

Michael Keith Butterfield

Dissertation presented for the degree of Doctor of Philosophy at the University of Stellenbosch. Promoter Dr F.C. Botha and Prof L. Warnich.

April 2007

DECLARATION

I, the undersigned, hereby declare that the work contained in this dissertation is my own original work and has not previously in its entirety or in part been submitted at any university for a degree. Where use was made of the work of others, it has been duly acknowledged in the text.

Signature:_____ Date: 18 October 2006

SUMMARY

Association analysis was used to improve the efficiency of breeding sugarcane varieties for the negatively correlated traits of resistance to sugarcane smut and the eldana stalk borer. 275 RFLP and 1056 AFLP markers were scored across a population of 77 genotypes representing the genetic variation present within the SASRI breeding programme. Genetic diversity analysis did not detect significant structure within the population. Regression analysis identified 64 markers significantly associated with smut rating and 115 markers associated with eldana rating at $r^2 > 6.25\%$. Individual markers with the largest effects explained 15.9% of the phenotypic variation in smut rating and 20.2% of the variation in eldana. Five markers were significantly associated with both smut and eldana. In each case the marker effect was negatively correlated between the two traits, suggesting that they are genetically as well as phenotypically negatively correlated.

Stepwise regression was used to identify sets of six markers explaining the maximum variation in phenotype. When the correlation between traits was not accounted for, the models predicted an increase in resistance in one trait, with an increase in susceptibility in the second trait. Accounting for the correlation resulted in models explaining 54% of the variation in smut resistance, and 62% of the variation in eldana resistance, with no undesirable correlated selection response in the second trait. Based on parent marker-type, cross combinations predicted to give more than 50% of progeny resistant to smut or eldana with no increase in susceptibility to the second trait were identified.

Association analysis was extended to the identification of groups of markers in linkage disequilibrium (LD) within the population. Methods of constructing whole-population LD maps were compared, and validated in an existing data set. In the validation population, 58% of haplotypes identified in LD were longer than 10 cM, confirming the potential of LD mapping as a tool in sugarcane breeding.

Using the method developed, a whole-population LD map consisting of 841 markers in 231 haplotypes was constructed for the mapping population of 77 SASRI genotypes, and an additional seven ancestral sugarcane clones important in the genealogy of modern germplasm. This is the first whole-population LD map constructed using association methods in sugarcane, or any other crop to date. Haplotypes associated with resistance or susceptibility to smut and eldana were identified. Comparing haplotypes present in the mapping and ancestral populations allowed the origin of important linkage groups to be traced, and indicated that disequilibrium due to population structure (SD) was also present. Some cases of SD involved co-segregation of a haplotype associated with smut or eldana resistance with a haplotype associated with susceptibility. Regression models developed to predict resistance to smut and eldana were extended to account for haplotype co-segregation due to SD. This resulted in an improvement in identifying cross combinations predicted to give progeny resistant to both smut and eldana. Priority can be given to making these crosses followed by within-family selection to identify progeny resistant to smut and eldana, in addition to having high cane yield and high sucrose content.

OPSOMMING

Die doeltreffendheid van die suikerriet telingsprogram vir eldana-stamboorder en suikerrietbrand weerstandbiedendheid is verbeter deur van genetiese merker-assosiasie gebruik te maak. Merkers is geïdentifiseer deur 275 RLFP (restriksie fragment lengte polimorfismes) en 1056 AFLP (ge-amplifiseerde fragment lengte polimorfismes) DNS fragmente te ontleed in 'n populasie van 77 genotipes, wat verteenwoordigend is van die genetiese variasie binne die SASRI telingsprogram. Diversiteitsanalises het aangetoon dat daar geen duidelike struktuur binne hierdie geselekteerde groep bestaan nie. Met behulp van regressie analises is 64 merkers vir suikerrietbrand- en 115 merkers vir eldana-stamboorder weerstandbiedendheid geïdentifiseer (r^2>6.25%). Individuele merkers het tot soveel as 15.9% van suikerrietbrand en 20.2% van eldana-stamboorder weerstandbiedendheid verklaar. Vyf van die genetiese merkers het betekenisvolle assosiasie met beide eldana-stamboorder en suikerrietbrand weerstandbiedendheid getoon. In alle gevalle was die merker assosiasie negatief gekorreleer tussen die eienskappe wat aantoon dat die twee eienskappe genotipies en fenotipies negatief gekorreleerd is.

Met behulp van stapgewyse regressie analise is stelle van ses merkers geselekteer wat die maksimum variasie van elke fenotipe verklaar. Wanneer die korrelasie tussen die fenotipes buite rekening gelaat word, verklaar die stelle merkers 'n toename in bestandheid vir die een fenotipe en afname in die ander. Wanneer die assosiasie tussen die twee fenotipes in ag geneem word verklaar die merkers 54% van die suikerrietbrand en 62% van die eldana-stamboorder weerstandbiedendheid. Gesimuleerde toetskruisings, gebaseer op die ouer se genetiese merker profiel, voorspel dat meer as 50% van die nageslag weerstand teen suikerrietbrand of eldana-stamboorder het sonder 'n toename in die vatbaarheid vir die ander eienskap.

Merkers in koppelingsdisekwilibrium (KD) is ook met behulp van assosiasie analise geïdentifiseer. Verskillende metodes om KD genetiese kaarte saam te stel is vergelyk en die akkuraatheid van die verskillende metodes is teen 'n reeds geykte datastel getoets. Ongeveer 58% van die haplotipes in die KD genetiese kaarte het meer as 10 cM oorspan wat die potensiaal van KD kartering as 'n metode om die doeltreffendheid van suikerrietteling te verbeter bevestig.

Hierdie ontwikkelde metode is gebruik om 'n KD kaart van 841 merkers in 231 haplotipes binne die populasie van 77 individuele genotipes en sewe voorouers, wat 'n belangrike komponent van die stamboom van die huidige suikerriet kiemplasma uitmaak, saam te stel. Hierdie is die eerste KD kaart van 'n genetiese populasie in suikerriet of enige ander gewas. Haplotipes wat met vatbaarheid en weerstandbiedendheid gekoppel is, is geïdentifiseer en die oorsprong van die haplotipes wat aan die twee fenotipes gekoppel is, kon deur die analise bepaal word. Hierdie resultate het ook aangetoon dat daar disekwilibrium as gevolg van populasie struktuur is. Gevalle van ko-segregering van haplotipes vir suikerrietbrand of eldana-stamboorder bestandheid met haplotipes vir vatbaarheid was teenwoordig. Regressie modelle wat ontwikkel is om eldana-stamboorder en suikerrietbrand

weerstandbiedendheid te voorspel is verbeter sodat haplotipe ko-segregering as gevolg van PS uitgeskakel kon word. Dit het gelei tot die identifikasie van kruisings wat nageslag met weerstandbiedendheid teen beide suikerrietbrand en eldana-stamboorder sal lewer. Voorkeur kan nou verleen word aan die uitvoer van kruisings, gevolg deur inter-familiële seleksies, wat nie alleenlik tot verhoogde weerstandbiedendheid sal lei nie, maar ook hoë opbrengs en suikerinhoud sal lewer.

AKNOWLEDGEMENTS

Firstly to my supervisors, Dr. Frikkie Botha and Prof. Louise Warnich for their advice, encouragement and suggestions, and for not complaining about close deadlines.

To Karl Nuss, the last Head of Plant Breeding at SASRI. Karl, you have been an inspiration, an example and a rock of support. Thanks for everything over the years. I wish I could have completed this while you were still here.

To friends at CIRAD; Angelique D'Hont, Jean Christophe Glaszmann, Jerome Pauquet and Louis Marie Raboin for looking after me at CIRAD, providing their data on the R570 and LD mapping populations and collaborating on LD mapping approaches in sugarcane. *Merci beaucoup, tout le monde.* Also for introducing me to the concept of double-blind-factorial wine tasting where treatments are randomly allocated to samples in each successive replication, confirming for me that stochasticism really is the fundamental universal force.

To the National Research Foundation and the French Foreign Ministry for funding my stay at CIRAD to generate the AFLP data for the study.

To Roy Parfitt and plant breeding staff at SASRI for giving me the space and time to complete this work.

Thanks to Stuart Rutherford, Julie Richards and Lucy Thokoane for the RFLP data that started me thinking about whole-population genome mapping, and to Barbara Huckett for encouragement and advice.

To Steph, for being there at the end, for all your support and for generally being fabulous.

Last but not least, to the makers of Autumn Harvest Crackling for the one-litre screw cap bottle at R14.85.... which eased the pain after long nights behind the keyboard.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

AA	association analysis
AFLP	amplified fragment length polymorphism
bp	base-pairs
BSA	bulk segregant analysis
cDNA	copy DNA
cM	centiMorgans
CERF	Centre d'Essai, de Recherche et de Formation
CIRAD	Centre de Cooperation Internationale en Recherche Agronomique pour le Développement
DArT	diversity arrays technology
DNA	deoxyribonucleic acid
EST	expressed sequence tag
GAP	gene/allele pyramiding
HD	homology disequilibrium
HG	homo(eo)logy group
IBD	identity by descent
ICSB	International Consortium for Sugarcane Biotechnology
ISSCT	International Society of Sugar Cane Technologists
kb	kilobase-pairs
LG	linkage group
LD	linkage disequilibrium
MAS	marker assisted selection
MRCA	most recent common ancestor
N-J	neighbour-joining
PCR	polymerase chain reaction
QTA	quantitative trait allele
QTL	quantitative trait locus
RAF	randomly amplified DNA fingerprints
RAPD	randomly amplified polymorphic DNA
RFLP	restriction fragment length polymorphism
sar	sum of adjacent recombination coefficients
SASRI	South African Sugarcane Research Institute
SCAR	sequence characterised amplified regions
SD	structure disequilibrium
SNP	single nucleotide polymorphism

SSH	suppression subtractive hybridization
SSR	simple sequence repeats
TDT	transmission-disequilibrium test

CHAPTER 1.

Introduction

1.1. The sugar industry in South Africa.

The South African sugar industry located within the provinces of KwaZulu-Natal, Mpumalanga and the Eastern Cape is an important contributor to both the national and provincial economies, generating an estimated direct income of R6 billion per annum. This is derived from an average annual production of 2.5 million tons of sugar milled from 22 million tons of sugarcane grown over 430 000ha of land. The sugar industry provides direct and indirect employment to approximately 350 000 people, and it is estimated that 1 million people are dependent on the industry for their livelihood (Anon, 2005a). Of the annual sugarcane crop, approximately 12% is produced by small-scale growers. An important initiative launched by the sugar industry in 2004 aims at ensuring a minimum 30% Black ownership of freehold sugarcane land by 2014 (Anon, 2005a), in line with the South African National Land Reform programme initiated in 1994 as part of the Reconstruction and Development Programme (Anon, 2005b).

For sustainable production, the sugar industry relies on high yielding sugarcane cultivars adapted to the different climatic zones where cane is grown, and resistant to the major pests and diseases prevalent in each region. The South African Sugar Association Experiment Station was established at Mt Edgecombe in 1925, with the specific task of testing and developing new cultivars suitable for local conditions (Nuss, 1998, Hewitt *et al.*, 2000). In 2005, the name of the organization was changed to the South African Sugarcane Research Institute, SASRI. Between 1945 and 2005, 48 sugarcane cultivars have been released by SASRI to the industry. Forty of these were bred and selected at SASRI, and eight were selected either from seed of crosses imported from India, or are foreign-bred varieties released after testing under local conditions. Of three important old or currently grown cultivars – *viz.* NCo310, NCo376 and N12 - Donovan (1996, 1998) estimated their benefit:cost ratio to the industry at 24.5:1, 65:1 and 8.6:1 respectively, which excludes the additional benefits of other technologies such as improved fertilization, weed control etc. This illustrates the importance of suitable cultivars in contributing to the economic sustainability of the South African sugar industry.

Four zones of production have been identified based on soil, climate and cane harvest cycle (Nuss, 1998), and the SASRI sugarcane variety improvement programme aims at developing cultivars suited to these regions. Breeding and selection is focused on the economic driver of sucrose yield per hectare; a product of cane yield per hectare and cane sucrose content. Resistance to pests and disease is, however, a major bottleneck in the development and release of new cultivars (Butterfield and Thomas, 1996), and breeding for resistance is a key aspect of the variety improvement programme. The main insect pest is the African sugarcane stalk borer *Eldana saccharina* Walker (Lepidoptera: Pyralidae) that causes extensive damage especially when cane is exposed to abiotic stress such as drought. As the larvae that cause the damage are protected within the stalk, chemical application is largely ineffective, and host-plant resistance has been the only method of control. The most serious viral diseases are caused by Sugarcane Mosaic Virus (SCMV), and Sugarcane Yellow Leaf Virus (SCYLV), while bacterial diseases include gumming (*Xanthomonas axonopodis* pv. *Vasculorum* (Cobb 1894) Vauterin *et al.* 1995), leaf scald (*Xanthomonas albilineans* (Ashby 1929) Dowson 1943) and ratoon stunting disease (*Leifsonia xyli* subsp. *xyli* (Davis *et al.*1984) Evtushenko 2000). Fungal diseases of importance are smut (*Ustilago scitaminea* Syd.), pokkah boeng (*Gibberella fujikuroi* (Sawada) Wollenw. and *G. subglutinans* (E.T. Edwards) P.E. Nelson, Tousson & Marasas), red rot (*Glomerella tucumanensis* (Speg.) Arx & E. Mull.) and brown rust (*Puccinia melanocephala* Syd. & P. Syd.). This long list of resistance traits for which strict thresholds are necessary in cultivar release, to which can be added the economic traits of yield and sucrose content, as well as agronomic traits such as erectness, resistance to lodging, good canopy development for weed control etc., affect the efficiency of cultivar development. In addition, the clonal nature of sugarcane cultivars and their vegetative mode of production, coupled with their complex genome, influence the scientific approach and structure of sugarcane breeding programmes.

1.2. The genome of sugarcane.

Sugarcane cultivars are advanced generation hybrids between two polyploid ancestor species, *S. officinarum* (2n = 80) and *S. spontaneum* (2n = 40 – 128) (Panje and Babu 1960, Price 1963). *S. officinarum* is an octoploid, having 80 chromosomes with a base chromosome number of x = 10, while *S. spontaneum* forms a polyploid series from 5-ploid to 16-ploid, with a base chromosome number of x = 8 (D'Hont *et al.*, 1998). *S. officinarum* is only found growing under cultivation in village garden in New Guinea, and has not been found in the wild. It is thought to have been selected by man from mutant forms of *S. robustum* (x = 10, 2n = 6x = 60), the predominant wild cane in New Guinea (Stevenson, 1965). *S. spontaneum* is a wild cane showing great phenotypic variation and with a wide

distribution through East Africa, parts of Eastern Europe, and most of Asia including the islands of Indonesia, the Philippines, Taiwan and Japan.

The original inter-specific hybridization events involved *S. officinarum* as the female parent, and *S. spontaneum* as the pollen donor. The cross is characterised by the functioning of unreduced female gametes, giving rise to progeny with a 2n+n chromosome complement. Backcrossing F1 hybrids to *S. officinarum* again results in 2n+n chromosome transmission, but in subsequent generations transmission reverts to normal (Bremer 1961). Structural differences between the genomes of the two species leads to meiotic instability and the production of aneuploid gametes (Burner and Legendre, 1993). Modern cultivars have complex polyploid-hybrid-aneuploid genomes with 100-130 chromosomes, of which 10% - 20% are of *S. spontaneum* origin, and the remainder from *S. officinarum* (D'Hont *et al.*, 1996). Within *S. officinarum* and *S. robustum*, preferential pairing between chromosomes within the same Homology group has been observed, whereas in general, crosses within *S. spontaneum* do not show preferential pairing. (Al-Janabi *et al.*, 1993; Al-Janabi *et al.*, 1994; Grivet *et al.*, 1996; Ming *et al.*, 1998; Hoarau *et al.*, 2001). In the commercial cultivar R570, however, chromosomes of *S. spontaneum* origin or recombinant chromosomes between *S. spontaneum* and *S. officinarum* do show preferential pairing (Grivet *et al.*, 1996; Hoarau *et al.*, 2001). Of all the major crop species, sugarcane has arguably the most complex genome (Grivet and Arruda, 2001), and this poses special challenges to breeding and genetic dissection.

1.3. Sugarcane breeding.

In addition to the complex genome, the fact that sugarcane cultivars are clones propagated vegetatively influences breeding strategies. Both additive and non-additive sources of genetic variation can be exploited, and breeding strategies need to take this into account. The high ploidy level, and the fact that meiotic instability results in sterility of many genotypes preclude the development and use of inbred lines in sugarcane. Worldwide, breeding programmes generally follow forms of recurrent selection (Jackson, 2005), testing large numbers of genotypes under high selection intensities to identify the few individuals with acceptable phenotypic values for the many traits required of a commercial cultivar. In the SASRI programme, typically, 250 000 new genotypes enter testing each year, derived from approximately 700 crosses between 300 parents. These candidates then follow a five-stage selection programme for the evaluation of yield, cane quality, agronomic and pest and disease resistance traits that takes between 11 to 15 years to complete. The number of genotypes under test, and the length of time of the testing period, is largely driven by the

numbers of traits required for selection, and the fact that single sugarcane clones occupy large areas of land for many years, and need to demonstrate stable performance over sites and seasons for yield and durable resistance under conditions of high disease pressure. Between 2000 and 2005, the average rate of cultivar release from SASRI has been approximately 1 per every 150 000 candidates tested.

The use of molecular tools – like DNA makers linked to quantitative trait loci and alleles (QTLs and QTAs) involved in phenotype of interest – has the potential to increase the efficiency of conventional breeding programmes. Markers are particularly effective to tag traits that are difficult or costly to measure, and traits that have low heritability (Lande and Thompson, 1990). The importance of pest and disease resistance traits in sugarcane, coupled with the fact that they can be difficult to measure reliably due to environmental variation (i.e. seasonal variation in pest populations and disease inoculum) makes these appropriate targets for marker assisted breeding. The theory for identification and use of markers for 'foreground' and 'background' selection in crops with inbred lines is well established (e.g. Hospital and Charcosset, 1997). Methods for out-breeding populations – both plant and animal – have also been developed (reviewed in Hoeschele *et al.*, 1997), but these generally assume a diploid genome, and also sometimes assume that prior information on mapped QTLs is available and that allelic relationships are known. The complex polyploid genome of sugarcane and the breeding strategies employed in cultivar development require that alternative methods for marker identification and utilization be developed in order to be effective. Conventional QTL identification experiments involve analyzing segregating progeny population derived from single crosses. Although these are effective for discovering and mapping loci of interest (QTLs), they can only identify the alleles (QTAs) present in the parents of the cross, and much of the genetic variation present in the breeding population will therefore remain undetected. This approach also requires establishing specific experiments to measure the phenotypes of interest. Unless these experiments are conducted over several sites and seasons, which adds significantly to the cost of data collection, the phenotypes measured may be subject to significant genotype by environment interactions, leading to the identification of QTAs that may not be robust under a wider range of conditions. In addition, in a complex polyploid genomic background, the presence or absence of single markers may be phenotypically uninformative, due to interactions between multiple alleles at the same locus.

Marker discovery through linkage disequilibrium or association analysis within germplasm populations offers an alternative approach to conventional QTL mapping in segregating progeny (Jannink *et al.*, 2001, Bresegello and Sorrels, 2006). Some of the advantages of

this approach are that a wide range of genetic variation existing within the population can be sampled, and the markers identified are less likely to be specific to a particular genetic background (Jannink *et al.*, 2001). Association analysis is beginning to be used in crops such as wheat, barley and potato (Bresegello and Sorrels 2005, Kraakman *et al.*, 2004, Simko *et al.*, 2004) but has not yet been rigorously tested in sugarcane.

1.4. Objectives of this study.

The aim of this study was to exploit features of the biology of sugarcane in designing strategies for marker discovery and use that could be applied directly in support of the conventional breeding programme, while containing costs by utilizing existing data collected routinely in the selection programme. The traits chosen for this purpose were resistance to sugarcane smut, and the eldana stalk borer. These are both among the more serious biotic challengers of sugarcane in South Africa, and had the additional advantage that some molecular characterization data were already available from a project on gene identification carried out at SASRI (Heinze *et al.*, 2001, Thokoane and Rutherford, 2001). More specifically, individual objectives of this study were:

- to use association analysis methods for identifying markers linked to smut and/or eldana resistance within a population representing the genetic variation present within the SASRI breeding germplasm.
- to develop methods of whole-population mapping to identify haplotypes present in linkage disequilibrium within a polyploid population.
- to validate the mapping methodology by using an independent data set with known linkage arrangement and determine if the extent of linkage disequilibrium in sugarcane populations is sufficient to be useful in molecular breeding.
- to create a whole-population map of the SASRI germplasm used for marker identification, and identify haplotype fragments containing smut and eldana resistance markers.
- to develop models for using markers and haplotypes associated with smut and/or eldana resistance as tools in an applied breeding programme.

Achieving these goals will provide new tools to improve the efficiency of developing superior sugarcane cultivars for the South African industry, and supply a framework for extending molecular breeding to additional phenotypes of importance.

1.5. References

Al-Janabi, SM, Honeycutt, RJ, McClelland M and Sobral, BWS. 1993. A genetic linkage map of *Saccharum spontaneum* L. 'SES 208'. Genetics 134: 1249-1260.

Al-Janabi, SM, Honeycutt, RJ and Sobral, BWS. 1994. Chromosome assortment in *Saccharum*. Theoretical and Applied Genetics 89: 959-963.

Anonymous. 2005a. South African Sugar Industry Directory 2005/2006. http://www.sugar.org.za.

Anonymous. 2005b. Sustainable land and agrarian reform: A contribution to vision 2014. Department of Land Affairs Strategic Plan 2005-2010. http://land.pwv.gov.za/publications /formal/ Strategic_Plans/Strategic%20planning%20for%202005%20-%202009%20FINAL %20DOCUMENT.pdf

Bremer, G. 1961. Problems in breeding and cytology of sugarcane II. The sugar cane breeding from a cytological viewpoint. Euphytica 10: 121-133.

Breseghello, F and Sorrels, ME. 2005. Association mapping of kernel size and milling quality in wheat. Genetics 172: 1165-1177.

Breseghello, F and Sorrels, ME. 2006. Association analysis as a strategy for improvement of quantitative traits in plants. Crop Science 46: 1323-1330.

Burner, DM and Legendre, BL. 1993. Chromosome transmission and meiotic stability of sugarcane (*Saccharum* spp.) hybrid derivatives. Crop Science 33: 600-606.

Butterfield, MK and Thomas, DW. 1996. Sucrose, yield and disease resistance characteristics of sugarcane varieties under test in the SASEX selection programme. Proceedings of the South African Sugar Technologists Association 70: 103-105.

D'Hont, A, Grivet, L, Feldmann, P, Rao, S, Berding, N and Glaszmann, JC. 1996. Characterisation of the double genome structure of modern sugarcane cultivars (*Saccharum* spp.) by molecular cytogenetics. Molecular and General Genetics 250: 405-413.

D'Hont, A, Ison, D, Alix, K, Roux, C and Glaszmann, JC. 1998. Determination of the basic chromosome numbers in the genus Saccharum by physical mapping of ribosomal RNA genes. Genome 41: 221-225.

Donovan, PA. 1996. An empirical evaluation of the sugarcane variety NCo310. Proceedings of the South African Sugar Technologists Association 70: 93-96.

Donovan, PA. 1998. The value of N12 in the Midlands North area. Proceedings of the South African Sugar Technologists Association 72: 35-41.

Grivet, L, D'Hont, A, Roques, D, Feldmann, P, Lanaud, C and Glaszmann, JC. 1996. RFLP mapping in cultivated sugarcane (*Saccharum* spp.): Genome organization in a highly polyploid and aneuploid interspecific hybrid. Genetics 142: 987-1000.

Grivet, L and Arruda, P. 2001. Sugarcane genomics: depicting the complex genome of an important tropical crop. Current Opinion in Plant Biology **5:** 122-127

Heinze, BS, Thokoane, LN, Williams, NJ, Barnes, JM and Rutherford, RS. 2001. The smut-sugarcane interaction as a model system for the integration of marker discovery and gene isolation. Proceedings of the South African Sugar Technologists Association 75:88-93.

Hewitt, PH, Donovan, PA and Carnegie, AJM. 2000. The South African Sugar Association Experiment Station at Mount Edgecombe – Milestones 1925-2000. Proceedings of the South African Sugar Technologists Association 74: 8-11.

Hoarau, JY, Offmann, B, D'Hont, A, Risterucci, AM, Roques, D, Glaszmann, JC and Grivet, L. 2001. Genetic dissection of a modern sugarcane cultivar (*Saccharum* spp).I. Genome mapping with AFLP markers. Theoretical and Applied Genetics 103: 84-97.

Hoeschele, I, Uimari, P, Grignola, FE, Zhang, Q and Gage, KM. 1997. Advances in statistical methods to map quantitative trait loci in outbred populations. Genetics 147: 1445-1457.

Hospital, F and Charcosset, A. 1997. Marker-assisted introgression of quantitative trait loci. Genetics 147:1469-1485.

Jackson P. 2005. Progress and prospects in genetic improvement in sucrose accumulation. Field Crops Research 92: 277-290.

Jannink, JL, Bink, MCAM and Jansen, RC. 2001. Using complex plant pedigrees to map valuable genes. Trends in Plant Science 6: 337-342.

Kraakman, ATW, Niks, RE, Van den Berg, PMMM, Stam, P and Van Eeuwijk, FA. 2004. Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars. Genetics 168: 435-446.

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

Ming, R, Liu, SC, Lin, YR, da Silva, J, Wilson, W, Braga, D, van Deynze, A, Wenslaff, TF, Wu, KK, Moore, PH, Burnquist, W, Sorrels, ME, Irvine, JE and Paterson, AH. 1998. Detailed alignment of *Saccharum* and *Sorghum* chromosomes: Comparative organization of closely related diploid and polyploid genomes. Genetics 150: 1663-1682.

Nuss, KJ. 1998. Aspects considered in the search for new farms for the Experiment Station. Proceedings of the South African Sugar Technologists Association 72: 42-45.

Panje, R and Babu, C. 1960. Studies in Saccharum spontaneum. Distribution and geographical association of chromosome numbers. Cytologia 25: 152-172.

Price, S. 1963. Cytogenetics of modern sugar canes. Economic botany 17: 97-105.

Simko, I, Haynes, KG and Jones, RW. 2004. Mining data from potato pedigrees: tracking the origin of susceptibility and resistance to *Verticillium dahliae* in North American cultivars through molecular marker analysis. Theoretical and Applied Genetics 108: 225-230.

Stevenson, GC. 1965. Genetics and Breeding of Sugarcane. Tropical Science Series. Longmans, Green and Co Ltd. London. 284pp.

Thokoane, LN and Rutherford, RS. 2001. cDNA-AFLP differential display of sugarcane (*Saccharum* spp. hybrids) genes induced by challenge with the fungal pathogen *Ustilago scitaminea* (sugarcane smut). Proceedings of the South African Sugar Technologists Association 75:104-107.

CHAPTER 2.

Literature review

2.1. Introduction.

The last two decades have seen rapid advances in molecular approaches in plant improvement that has lead to their routine adoption in the breeding programmes of some crops. This is perhaps reflected in the fact that a dedicated journal entitled 'Molecular Breeding' has been established, that published its first issue in March 1995 and is currently in its 18th volume. The term 'molecular breeding' is broad, and can include technologies such as the incorporation of novel genes into plants by transgenesis, or silencing endogenous genes through the introduction of specific DNA or RNA sequences. Here, and in chapters to follow, 'molecular breeding' will be used in a narrower sense to describe methods using measures of genetic polymorphism within populations to provide breeders with additional information that can be used to improve breeding and selection decisions.

The basic premise behind molecular breeding is simple: if variation in phenotype for a trait is caused by genetic polymorphism at one or several gene loci, then direct measurement of the genetic polymorphisms will allow the phenotype to be predicted at some level of confidence. It may not be possible to detect the exact causal genetic locus, but if polymorphism can be detected at a linked locus, the DNA marker linked to the gene can be used as a surrogate for predicting the phenotype. Technologies for generating many types of DNA markers are now available, and will not be reviewed here. In addition, many different statistical techniques have been developed in order to identify marker-trait associations. These have been reviewed and described in standard texts such as Lynch and Walsh (1998), Liu, (1998) and Balding *et al.* (2003). Different marker identification methods are appropriate under different circumstances, and choice depends on factors such as mating system (e.g. inbreeding versus outcrossing), breeding strategy (e.g. selfing, backcrossing, cross breeding or hybrid breeding), and genome structure (e.g. diploid versus polyploid). Broadly speaking, all strategies to identify marker-trait association fall into two classes: those using a gene mapping approach within a population of segregating progeny from a bi-parental cross, or those using an association analysis (AA) approach involving a range of different population types. Association analysis is also commonly referred to as linkage disequilibrium (LD) analysis. It is important to note, however, that what is common between the two types of approach is that they both rely on the presence of linkage disequilibrium in order to detect the association (Lynch and Walsh, 1998).

The aim of work reported in the chapters to follow is to use an association analysis or linkage disequilibrium approach to identify markers for pest and disease resistance in sugarcane, and develop strategies for their application in breeding. As such, gene mapping approaches used in other plant crops are not directly relevant, and will not be reviewed here. Mapping studies in sugarcane will be reviewed, however. These describe the development of genetic analysis in sugarcane, and constitute the body of knowledge upon which the rationale for this study is based. Association analysis approaches in other crop species will be reviewed, as these are directly relevant to the objectives and execution of this study. This will place the current study within a broader context, and highlight the contribution that this study makes to association analysis for plant improvement in general, and for sugarcane in particular.

The objective of this review chapter will be firstly to provide a brief description of the historical development of marker-trait association and molecular breeding approaches in general. This will be followed by a description of sugarcane genetic mapping work completed to date, and a summary of marker-trait associations detected within sugarcane mapping populations. The review will conclude with an examination of linkage disequilibrium approaches developed in human systems and how these have been applied in different crop species to date.

2.2. A brief history of the identification and application of genetic markers.

The concept of using genetic association between characters to explain phenotypic variation is not new. In 1918, Fernandus Payne conducted a series of elegant breeding experiments with *Drosophila*. He had evidence to suggest that variation in scutellum bristle number was a sex-linked trait, and by crossing selected individuals in such a way that the x-chromosome carrying the mutation was always inherited in the breeding line, was able to increase the number of scutellum bristles in individuals from 4 up to 15 (Payne, 1918). Knowledge of the position of the gene and a crossing strategy to exploit this information therefore resulted in an increase in efficiency for breeding for high bristle number in *Drosophila*.

Karl Sax (1923), was interested in size differences between bean genotypes, and hypothesized that if a linked qualitative factor could be identified, then the 'Mendelizing factors' governing bean size variation could be studied. Within the germplasm he worked with, he found a strong association between seed coat colour and bean size, which enabled him to select large seeded progeny based on their segregation for seed coat colour. Other early studies on the use of correlated characters to study variation in quantitative traits were done in tomato (Lindstrom, 1924), peas (Rasmusson, 1927), maize (Lindstrom, 1931) and

barley (Wexelsen, 1934). Although these studies were concerned with the association between phenotypic traits, the basic premise of using a correlated response from genetically linked characters is fundamentally the same as the current approach of using molecular markers linked to phenotype.

Morphological markers proved to be limited in their application, and it wasn't until genetic markers became available that analysis of trait association began to progress. Liu (1998) described the 'evolution' of genetic markers as beginning with allozymes and isozymes in the pre-recombinant DNA era, and moving to restriction fragment length polymorphisms (RFLPs) in the pre-PCR (polymerase chain reaction) era. With the development of PCR technology, markers entered the Oligo-era with the advent of systems such as randomly amplified polymorphic DNAs (RAPDs), microsatellites, and amplified fragment length polymorphisms (AFLPs), finally entering the Cyber-genetics era as DNA sequence information, such as single nucleotide polymorphisms (SNPs). Development of these technologies will not be reviewed, but early application of markers in sugarcane in the pre-recombinant DNA era will be described.

Isoenzymes had been used as molecular markers for sugarcane variety identification as early as 1969 (Heinz, 1969). The first attempt to use molecular markers in sugarcane breeding was described by Roughan *et al.* (1971). In this study, β-amylase isoenzyme variation in *S. officinarum*, *S. spontaneum* and F1 hybrid progeny was studied in an attempt to elucidate the nature of stalk starch inheritance. The authors found no obvious correlation between different β-amylase isoenzyme bands and stalk starch levels in individual genotypes, but showed that isoenzyme bands could clearly differentiate between *S. spontaneum* and *S. officinarum* genotypes as well as between true hybrid progeny and progeny derived from self-fertilization. By comparing the isoenzyme bands present in the *S. officinarum* parents and the hybrid progeny, they were also able to conclude that the 2n gametes contributed by the *S. officinarum* in hybrids with *S. spontaneum* resulted from an event following the first reduction division of meiosis, and were not due to the formation of unreduced egg-cells. The mechanism of 2n gamete formation in *S. officinarum* had previously been the subject of much debate, and had been described by Stevenson (1965) as "...one of the most absorbing problems in sugar cane cytogenetics." Thus although Roughan *et al.* were not able to use isoenzyme markers to breed or select for progeny with low starch content in the stalk, their results were able to shed some light on issues of genetic diversity within *Saccharum*, as well as issues of basic genetics and cytology. This has been a feature of much of the subsequent work published on sugarcane molecular markers and mapping to be reviewed here. Until recently, molecular analysis of sugarcane and its

relatives has focused on developing a better understanding of genome structure and organization, rather than developing direct applications for breeding. This has been invaluable, as until the mid 1990's basic knowledge such as the base chromosome number of *Saccharum* species, and chromosome pairing behavior in pure species and commercial hybrids was lacking.

From the initial work of Roughan *et al*, it was another 18 years before the subject of molecular markers in sugarcane received serious attention. Glaszmann *et al.* (1989) surveyed isoenzyme variation for nine enzymes among 39 genotypes of *S. officinarum*, *S. robustum*, *S. spontaneum* and *Erianthus*. Multivariate analysis showed that *S. officinarum* and *S. robustum* clustered together, supporting the hypothesis that *S. officinarum* is related to - and presumably derived from - *S. robustum* (Roach and Daniels, 1987). *S. spontaneum* clustered separately, as did the single genotype of *Erianthus*. The multiple isoenzyme bands of unequal intensities observed suggested a high degree of allelic variation and gene dosage, consistent with the high ploidy level and heterozygosity of *Saccharum*. Analysis of selfed progeny of *S. spontaneum* and of the commercial cultivar R570, however, showed some examples of monogenic – i.e. single-dose – segregation, and Glaszmann *et al.* (1989) remarked that these may be useful in investigating patterns of inheritance in sugarcane. The key publication by Wu *et al.* (1992) describing a method for mapping single-dose and double-dose DNA markers in segregating populations of polyploids confirmed the suggestion by Glaszmann *et al.* (1989) on the utility of monogenic markers for genetic analysis. Wu *et al.* (1992), showed that single-dose markers – i.e. present in one copy in parent 1 and absent in parent 2 - segregate 1:1 in progeny of a cross, and can be mapped using conventional theory developed for diploid species. This, along with advances in molecular methods such as the development of polymerase chain reaction (PCR) technology, laid the foundation for greatly increased activity in sugarcane molecular marker studies in the 1990's.

2.3. Genome mapping in sugarcane.

The first genetic map for *Saccharum* was published by Al-Janabi *et al.* (1993), for the *S. spontaneum* clone 'SES 208' (2n = 64). The mapping population used consisted of 88 progeny derived from a cross between SES 208, and a doubled-haploid of SES 208 produced through anther culture. Randomly amplified polymorphic DNA (RAPD) primers were used to generate markers. Of 279 scorable polymorphisms, 208 were single-dose, and 176 of these could be allocated to 41 linkage groups. Segregation analysis showed no preferential pairing between chromosomes, suggesting that *S. spontaneum* SES 208 behaves meiotically like an autopolyploid. Although the results were not definitive,

segregation was consistent with an auto-octaploid genome. This was a significant finding, as at this time the base chromosome number of the different *Saccharum* species was still unknown. This map was extended by Da Silva *et al.* (1993, 1995) by the addition of 276 restriction fragment length polymorphisms (RFLPs), resulting in a map of 64 linkage groups assigned to 8 Homology Groups, providing further evidence that SES 208 is an auto-octaploid, with a basic chromosome number of x=8.

In 1996, a genetic map of the *S. officinarum* clone 'La Purple' was published by Mudge *et al.* (1996), constructed from the analysis of 84 F1 progeny derived from a cross between the *S. officinarum* clone La Purple and the *S. robustum* clone 'Molokai 5829'. This map consisted of 160 single-dose RAPD markers and one morphological marker, organized into 51 linkage groups. Of significance in this study was the detection of 10.9% of linkages in repulsion, suggesting preferential pairing between some homologous chromosomes, and the possibility that *S. officinarum* is a segmental allopolyploid on an evolutionary path to diploidization. This is in contrast to the results of Al-Janabi *et al.* (1993) for *S. spontaneum*. In addition, linkage of a RADP marker to a putative dominant gene for susceptibility to eyespot disease specific to the *S. officinarum* parent was detected, suggesting that marker-assisted identification of eyespot susceptibility may be possible.

The first comprehensive map of a commercial hybrid sugarcane cultivar was published in 1996 by Grivet *et al*, for the commercial cultivar R570, bred at the Centre d'Essai, de Recherche et de Formation (CERF) in Reunion. Some preliminary mapping work on this cultivar had been done (D'Hont *et al.* 1994 and Grivet *et al.* 1994), but the number of progeny used in these studies was to small to order most of the markers. The map of Grivet *et al.* (1996) was constructed from the analysis of RFLP marker segregation in 77 progeny derived from self-pollination of R570, and genotypes of *S. officinarum*, *S. spontaneum* and *S. barberi* were included in RFLP screening in order to trace the species origin of individual markers in the commercial hybrid. The resulting map consisted of 408 markers assembled into 96 linkage groups, which were tentatively assigned to 10 Homology Groups on the basis of shared probes. Of the mapped markers, 73 were of putative *S. officinarum* origin, and 63 were derived from *S. spontaneum*. Of particular interest was the detection of at least six recombination events between chromosomes of the two ancestral species within the genome of R570, as it had previously been assumed that interspecific recombination did not occur within hybrid sugarcane (Price 1967, Berding and Roach, 1987). In addition strong preferential pairing was observed between chromosomes of *S. spontaneum* origin, and also between a *S. spontaneum* and a putative recombinant chromosome. Thus although preferential pairing does not occur within *S. spontaneum* (Al-Janabi *et al.*, 1994), within the

hybrid genome that contains 10 to 15% of the *S. spontaneum* genome (predicted by Simmonds, 1976, demonstrated by D'Hont *et al*, 1996, 1998), chromosomes of *S. spontaneum* do pair preferentially, rather than with their *S. officinarum*-derived homoeologues. A third point of importance in Grivet *et al.* (1996) was that by including ancestral clones in the pedigree of R570 in the marker analysis, it could be established that the reported parents of R570, *viz.* H32-8560 and R445 were in fact the true parents of R570, but that identity of the reported grandparents could be disputed on molecular evidence. This illustrates one of the important applications of molecular markers in breeding, as identification of the correct pedigree of cultivars can be as important in breeding decisions as tracing the inheritance of important characters from parents to offspring.

The map of R570 was later extended by Hoarau *et al.* (2001), by adding 887 amplified fragment length polymorphism (AFLP) markers, which were distributed across 120 linkage groups (LGs). By examining species-specific markers, 11 LGs could be assigned a *S. spontaneum* origin, and 79 a *S. officinarum* origin, with 11 LGs showing recombination between ancestral chromosomes. Thirteen linkage group pairs showed repulsion phase linkages, implying a high degree of preferential pairing. The total length of the map was 5849 cM, which is estimated to be about one-third of the predicted genome size. Although this represents a fairly sparse coverage, this map contains the most markers of any individual *Saccharum* map published to date.

Knowledge of genome architecture of *Saccharum* and its relatives inferred through molecular markers and mapping was advanced through the work of Guimaraes *et al.* (1997). They used RFLP probes previously used in sorghum, maize and *S. spontaneum* (Da Silva *et al.*, 1995) to map La Purple (*S. officinarum)* and Molokai 5829 (*S. robustum).* These two genotypes had previously been mapped with RADPs by Mudge *et al.* (1996). The rationale for comparative mapping with sorghum is based on the fact that there are no known wild *Saccharum* diploids, and sorghum is thought to be the closest diploid relative as it shares a base chromosome number of x = 10 with *S. officinarum* and *S. robustum.* Results showed no changes in marker order between sorghum and *S. officinarum*, and only 4 inversions between sorghum and *S. robustum* – two of which might be artifacts of the relatively small population size used for mapping. The strong colinearity between *Saccharum* and sorghum confirms the potential utility of sorghum as a model system for the dissection of the more complex genome of *Saccharum.*

The exploitation of comparative mapping was taken a step further by Ming *et al.* (1998) with the detailed alignment of maps of four *Saccharum* clones with an established high-density

map of sorghum. The genotypes used were two *S. spontaneum* genotypes (IND 81-146 and PIN 84-1), and two putative *S. officinarum* clones (Green German and Muntok Java). *The S. officinarum* designation of the Green German and Muntok Java clones used is however in dispute, and these may in fact be interspecific hybrids. Map results confirmed the high degree of colinearity between *Saccharum* and sorghum, with at least five *Saccharum* Homologous Groups corresponding almost completely to single chromosomes of sorghum. Some regions of extensive intrachromosomal rearrangements were however also detected, for example between sorghum linkage group C, and sugarcane Homology Groups (HG) 3 and 8. *S. spontaneum* and *S. officinarum* were distinguished from sorghum by only one inter- and two intra-chromosomal rearrangements, whereas 11 rearrangements differentiated *S. spontaneum* and *S. officinarum*. The higher degree of restructuring between *S. spontaneum* and *S. officinarum* is consistent with the difference in their base chromosome numbers of x = 8 and x = 10 respectively. The observation that HGs 6 and 7 in *S. spontaneum* co-locate to two sorghum linkage groups (LG), whereas the same HG in *S. officinarum* corresponds to one sorghum LG suggests that fusion of ancestral chromosomes is one possible explanation for this difference in base number. Differences in pairing behavior between the species was re-confirmed, with *S. robustum* showing higher levels of preferential pairing than *S. officinarum*, and complete random assortment in *S. spontaneum*. A perhaps surprising finding was that across comparable genetic regions, the rate of recombination of *Saccharum* was threefold higher than that of sorghum, despite their base genome size being similar. Using the same mapping populations, with the addition of new markers derived from a further set of 12 RFLP probes, Ming *et al.* (2002a) constructed a consensus map of *Saccharum* from the maps of the four parent genotypes. This map corresponds to only 70% of the sorghum map, illustrating that the *Saccharum* map remains incomplete, with gaps on most, if not all chromosomes. Maps of the individual genotypes ranged from 1395 cM (PIN 84-1), to 2466 cM (Green German) in size, which is considerable less than that of the 5849 cM R570 map. Averaged across the four *Saccharum* parents, 36% of single dose markers remained unmapped, and the authors note that the proportion of the unmapped genome may be higher than the proportion of unmapped markers. This is possibly due to the ploidy level of sugarcane, and the fact that only single-dose markers can be reliably mapped.

The most recent genetic map for sugarcane was published by Aitken *et al.* (2005). The population used was a cross between the elite cultivar Q165, and a *S. officinarum* clone IJ76-514, and a map of the Q165 parent was constructed. 1075 AFLP, randomly amplified DNA fingerprints (RAF) and simple sequence repeats (SSR) markers were mapped into 136 linkage groups. Repulsion phase linkage detected preferential pairing for 40 LGs, which

formed 11 LG pairs and three multi-chromosome pairing groups. Using SSRs, double-dose markers and repulsion phase linkages, 126 of the LGs were assigned to eight Homology Groups. Two HGs were each represented by two sets of LGs. These sets of LGs potentially correspond to *S. officinarum* chromosomes, with each set aligning to either end of one or two larger LGs. The larger chromosomes in the two HGs potentially correspond to *S. spontaneum* chromosomes. This suggestion is consistent with the different basic chromosome number of the two species that are hybridised to form sugarcane cultivars, *S. spontaneum* (x=8) and *S. officinarum* (x=10), and illustrates the structural relationship between the genomes of these two species. Although at 9058 cM, this map is 'longer' than the R570 map, the Q165 map contains fewer markers.

The maps described above were all derived from conventional mapping methodologies of analyzing single-dose polymorphisms in progeny populations derived from bi-parental crosses or self-fertilizations. The maps have all contributed greatly to our understanding of the *Saccharum* genome, the differences between *Saccharum* species, and how these differences contribute to the genome complexity of hybrid cultivars. These basic mapping studies have also resulted in the identification of markers linked to putative QTLs for several traits. So far, the maps have not been applied in breeding programmes in the countries in which they have been developed, but nevertheless they will provide an invaluable resource for future work if targeting specific areas of interest in the genome is required.

2.4. Identifying marker-trait associations in sugarcane by mapping.

The first molecular marker in sugarcane linked to phenotype identified through a mapping approach was published by Daugrois *et al.* (1996). In the selfed progeny population of R570 used for map construction described above (Grivet *et al.*, 1996), a clear 3:1 segregation was observed for resistance to brown rust, suggesting the presence of a single-copy dominant resistance gene. This putative resistance gene could be placed on the R570 map 10 cM from an RFLP marker derived from the probe CDSR29. Further targeted mapping using selective genotyping and AFLP markers was able to identify new markers 1.9 cM and 2.2 cM on either side of the rust resistance gene (Asnaghi *et al.*, 2004). Mapping of markers linked to phenotype in R570 was extended by Hoarau *et al.* (2002), who investigated QTLs for sugar yield components, *viz.*, brix, stalk length, stalk diameter and stalk number. These traits were measured in 295 selfed progeny in two crop seasons; plant cane and first ratoon. A total of 40 putative QTLs were identified for the four traits, but 35 of these were specific to one season; only 5 QTLs were detected across both years, at $p = 0.005$. Individual markers ascribed between three and seven percent of the phenotypic variation observed for the

different traits. Multiple regression models ascribed between 30% (stalk diameter in plant cane), and 55% (brix in first ratoon) of the phenotypic variation in the yield component traits. The authors also examined the direction of the marker effect in relation to its origin. Of markers located on 16 linkage groups whose species origin could be determined, 11 showed an effect in the predicted direction; e.g. positive effect on brix for *S. officinarum* specific markers, or positive effect on stalk number for *S. spontaneum* derived markers. In relation to the five markers observed with the opposite effect to that expected, it should be noted that in other crops such as tomato, QTLs associated with favourable effects have been identified in parents with unfavourable phenotype (e.g. Tanksley *et al.*, 1996).

The other *Saccharum* maps described in Chapter 2.2. have also been used to identify QTLs for a range of traits. Ming *et al.* (2001) used the mapping populations previously described (Ming *et al.* 1998, and Ming *et al.* 2002a) to examine QTLs for sugar content. Full multiple regression models explained 65% and 68% of the phenotypic variation in sucrose content observed in the two populations. However these models contained 14 and 22 individual QTLs respectively and are likely to be statistically over-fitted, as this number of markers correspond to 16384 (2^{14}) and 4194304 (2^{22}) different genotype classes. The independent identification of QTLs in the same genic region from *S. officinarum* and *S. spontaneum* strongly suggest that these regions are important in the control of sucrose content. The discovery of some QTLs with positive effects from the low-sucrose wild genotypes indicated the potential of using markers to introgress desirable new genes into breeding populations from wild germplasm. Some sugarcane QTLs mapped to equivalent regions in maize containing the key sugar metabolizing enzymes ADP glucose pyrophosphorylase and sucrose phosphate synthase. The effect of QTL dose on phenotype was investigated for 4 markers whose dosage could be determined (zero, one or two copies). Graphical representation of the data suggested a less-than-additive effect, although in each case a linear model gave a significantly better fit than a quadratic or cubic model.

Ming *et al.* (2002b) extended the study to include QTL discovery for sugar yield and yield component traits (pol, stalk weight, stalk number, and ash and fibre content). For the six traits, 102 QTLs were identified, of which 61 were mapped, with 50 of them clustering in 12 genomic regions on seven Homology Groups. Individual clusters contained between two and nine QTLs, with the possibility that some of these correspond to different alleles at the same locus. The percentage variance accounted for by individual QTLs ranged from 3.8% to 16.2%. Full regression models for each trait explained from 71.6% of the variation for stalk weight, to 6.1% for stalk number. These models, however, contain 14 and one QTL respectively, accounting for the large discrepancy in the amount of trait variation explained.

The allele effects of most of the QTLs were consistent with the phenotype of the parent from which they were derived, apart from a few exceptions where progeny exhibited transgressive segregation. This re-enforces the finding of Ming *et al.* (2002a), and suggests that the use of marker information in breeding could allow the identification of new gene combinations for specific traits.

The utility of sorghum as a model system for sugarcane has been demonstrated through QTL identification. In the comparative mapping study of Guimaraes *et al.* (1997), a *S. officinarum* marker strongly associated with short-day flowering was identified. The marker is located in a linkage group that is homoeologous to regions of the genomes of sorghum, rice and maize that also contain QTLs associated with flowering traits, suggesting that comparative mapping could be used to study other traits of interest. Ming *et al.* (2002c) compared QTLs identified for flowering time and plant height with QTLs for the same traits in sorghum. For plant height, QTLs identified in sugarcane corresponded to four of six QTLs mapped in sorghum. Flowering QTLs in sugarcane corresponded to one of three QTLs known in sorghum. Chromosomal rearrangements between sugarcane and sorghum were evident, but despite this it seems that the genomes share a high degree of colinearity that can be exploited through comparative mapping. This was confirmed by Jordan *et al.* (2004) who found that RFLP markers associated with tillering in a sugarcane bi-parental cross were located within or near QTLs for tillering that had been mapped in sorghum.

A recent publication by Reffay *et al.* (2005) has aimed at combining a mapping approach with pedigree analysis in order to identify genome regions of commercial interest. The *S. spontaneum* clone 'Mandalay' has been an important contributor in the Australian breeding programme, and occurs within the pedigrees of 25 recently released cultivars, as well as valuable unreleased parental germplasm. Mandalay is a grandparent of the genotype MQ77-340, whose genome has been mapped from a bi-parental cross with the commercial cultivar Q117, with the objective of identifying chromosome regions from Mandalay associated with sucrose and fibre content. Of 352 markers generated, 86 could be identified as of Mandalay descent, and 64 of these were placed on the MQ77-340 map that was constructed. This map comprises 3600 cM, accounting for an estimated 20% of the MQ77-340 genome. For three sucrose related traits, *viz.* pol, brix and commercial cane sugar (CCS), 23 marker-trait associations were identified across two seasons. These traits exhibit a strong auto-correlation, as brix is essentially a measure of pol plus additional soluble solids, and CCS is an index of pol, brix and fibre. One marker was associated with all three traits over both seasons. A further 47 markers were identified for the traits of fibre content and stalk weight, and the composite traits of cane yield and sugar yield. For all traits studied,

marker association explained 3-7% of the phenotypic variation within the mapping population, with some of the Mandalay specific markers having an effect opposite to that expected. The authors conclude that the study has allowed the identification of genic regions from Mandalay retained through selection in important breeding germplasm, and that this will assist in using markers for breeding in the future.

A more recent report by Aitken *et al.* (2006) described the identification of markers linked to sucrose accumulation in a mapping population derived from cross between a commercial sugarcane cultivar and a *S. officinarum* clone referred to in Chapter 2.3. (Aitken *et al.*, 2005). In the mapping population, 37 marker associations were detected for brix and sucrose content. Of these 37 putative QTLs, 30 were clustered into 12 genomic regions in six of the eight Homology Groups. Each QTL explained from 3 to 9% of the phenotypic variation observed. Markers associated with either an increase in brix/sucrose or a decrease in brix/sucrose were identified on linkage groups belonging to the same Homology Group, suggesting that a number of the markers were allelic forms of the same genes.

The markers identified in the mapping studies described above have not been applied in breeding programmes for a number of reasons. Although R570 is a commercial cultivar that has been used as a parent in breeding programmes, the markers identified for the yield component traits have not yet been verified in the broader breeding population, and so far have not been used in breeding (Raboin, personal communication[1].). In the various studies reported by Ming, the parental genotypes used for mapping are not clones that are generally used in breeding programmes, and the markers identified have not been tested in commercial germplasm. In addition, as the majority of the F1 progeny of one of the crosses do not flower (Ming *et al.*, 2002c), this poses an obvious impediment to breeding. The sucrose related QTLs reported by Aitken *et al.* (2006) are intended for use in the Australian sugarcane breeding programme, but no reports on their use or validation have appeared in the literature to date. The approach of analyzing segregating progeny populations has been very effective for QTL discovery and mapping in many crop species, including sugarcane, as described above. As outlined briefly in Chapter 1.3 and above, however, this approach often is difficult to translate directly into breeding applications. Marker identification and gene discovery through association and linkage disequilibrium (LD) methods within more complex populations has some benefits over traditional methods, and will be discussed in the remainder of this review chapter.

[1] Louis Marie Raboin: CIRAD. Currently Rice Breeder at CIRAD, Madagascar. email raboin@cirad.fr.

2.5. Development of linkage disequilibrium or association analysis methods.

Association and linkage disequilibrium methods have been pioneered in human studies (e.g. Lander and Schork, 1994; Risch and Merikangas, 1996), for the identification of genes causing specific diseases. Because conventional mapping populations cannot be created in humans, alternative methods have been required. This had lead to the development of approaches and methodologies to exploit linkage disequilibrium and population structure to map genes of interest in humans. Reviews of these methods can be found in Reich *et al.* (2001), Pritchard and Przeworski (2001) Goldstein and Weale (2001) and Palmer and Cardon (2005).

Collins and Morton (1998), defined allelic association – synonymous with linkage disequilibrium – as the dependence of allele frequencies at two loci, with the natural measure of the coefficient of association, ρ, being; $\rho_{ij} = (1 - \theta_{ij})^t \sim \exp(-t\theta_{ij})$; where θ_{ij} is the recombination frequency between loci I and J, and t is the number of generations. They described the high-resolution mapping of a disease gene in humans using an existing high-density marker map. In 2001, Morton *et al.* compared ρ against seven other statistical measures of allelic association including covariance, correlation, regression and frequency difference, and found that the best of them was only 80% as efficient as ρ. The coefficient of association also had the advantage of being less sensitive to variations in marker allele frequency than the alternative measures.

Since the early advocacy of association-based methods it has been recognized that population substructure or admixture will result in the 'detection' of spurious associations (i.e. type 1 errors) when there are allele frequency differences between the population sub-groups (e.g. Lander and Schork, 1994, Ewens and Spielman, 1995). One method developed to control stratification is the transmission-disequilibrium test (TDT) or case-control test (Spielman *et al.*, 1993; Ewens and Spielman, 1995). This design requires the presence of an affected individual and both parents, and the analysis of the alleles transmitted to the affected offspring (the case) versus those not transmitted (the control). As each case-control pair is matched within a family, allele frequency differences at the population level become irrelevant and the problem of stratification disappears. Although they have been widely used in human disease studies, TDT design are logistically complicated due to the requirement of parent-offspring triads, and may also result in reduced power to detect genetic associations (Cardon and Palmer, 2003).

In order to avoid the logistics of a TDT approach to controlling population stratification, Pritchard and Rosenberg (1999) proposed a method that uses information from unlinked markers to test for stratification. By analyzing the allele frequency at independent loci scattered through the genome, the null hypothesis that allele frequencies are the same for the case and control groups can be tested. If the test shows that no stratification is present, analysis of case-control data can proceed without the risk of detecting false associations. This approach was extended by Pritchard *et al.* (2000a; 2000b) to account for population structure in association mapping, and implemented in the software programs Structure 2.1 and STRAT (available from http://pritch.bsd.uchicago.edu/software.html. Verified 1 October 2006). In this method, if population structure is detected, individuals are assigned to the appropriate subpopulation and analysis proceeds within each group, effectively reducing the type 1 error rate and allowing the detection of different associations in different populations, should they exist. In simulation studies with populations of different degrees of admixture, the STRAT method provided equal statistical power to that of the TDT method, and outperformed TDT in situations where there were conflicting marker associations in different groups. This method is becoming increasingly used in association studies in plants (e.g. Thornsberry *et al.* 2001, Kraakman *et al.* 2004, Breseghello and Sorrels, 2005). Thornsberry *et al.* (2001) reported that accounting for population structure reduced the number of false positive associations detected for flowering-time variation in maize by 80%. It should be noted, however, that the method pre-supposes some map information, as it uses markers known to be unlinked in order to detect population stratification. In addition, the method assumed co-dominant markers with known allelic relationships. Dominant markers can be used if each marker is treated as a haploid allele with missing data, and this 'fix' is valid under the assumption of population structure without admixture; i.e. each individual comes purely from one of K subpopulations. If admixture is present, this fix is not valid. Estimates are likely to be unbiased, however, if large numbers of loci are used. (Pritchard and Wen, 2003).

The success of the linkage disequilibrium and association analysis methods in humans has stimulated interest in using similar approaches in plants. In the past 10 years, linkage disequilibrium studies and association analysis has been conducted in some major crops, and will be discussed in the following section.

2.6. Linkage disequilibrium and association analysis in plants.

The benefits of association and LD methods in plants were reviewed by Jannink *et al.* (2001), who listed the following advantages over bi-parental population approaches;

- Sampling of the full allelic variation within the breeding population, as opposed to that occurring only within a bi-parental progeny population;
- Detecting QTL effects within a diverse genetic background representing elite varieties means that markers detected are less likely to be background-specific, and more widely applicable in breeding;
- Using material under routine evaluation reduces the cost of collecting phenotypic data, as special experiments are not required;
- The ability to use retrospective analysis on old genotypes across several generations. Here the authors mention the ability to save old genotypes as seed, but in the case of a vegetatively propagated crop such as sugarcane, individual genotypes can be maintained within clonal collections over many years, providing a valuable resource for analyzing DNA variation and phenotype across generations.

Jannink *et al.* (2001) define association methods as relying on unrecorded sources of disequilibrium, including sources such as admixture, which cause transient disequilibrium in the absence of actual linkage, and define linkage methods as relying on family relationships to estimate the probability of chromosome segments within a pedigree of being identical by descent (IBD). Other authors (e.g. Collins and Morton, 1998) regard association and linkage disequilibrium to be synonymous. In this thesis, the definitions of Jannink *et al.* (2001) will be followed, in order to differentiate between physical linkage and other possible causes of association.

Flint-Garcia *et al.* (2003) gave a general overview of current knowledge of LD in plants, and the potential of association analysis to investigate genetic polymorphisms associated with traits at the population level. They discuss the work of Tenaillon *et al.* (2001), Remington *et al.* (2001) and Rafalski (2002) in maize, where the extent of LD has been found to vary considerably with the population studied. In diverse germplasm, Tenaillon *et al.* (2001) found that LD did not extend beyond 200 bp for 21 loci on chromosome 1. In contrast, in elite maize populations Rafalski (2002) and Ching *et al.* (2002) reported that LD extends over 100 kb for the *adh1* and *y1* loci, and decay in LD was not detectable over a 300-500 bp range for 18 other genes. The situation in *Arabidopsis* is quite different, with LD extending over longer segments of the genome. Hagenblad and Nordborg (2002) found that LD extended up to 250 kb, equivalent to about 1 cM of recombination, in 14 sequenced fragments in the region

of the *FRIGIDA* flowering locus. Similarly, Nordborg *et al.* (2002) found that LD in the region of 250 kb for 163 genome-wide SNP loci across 76 *Arabidopsis* accessions. The difference in the extent of linkage disequilibrium between maize and *Arabidopsis* is not unexpected. In self-fertilizing species (e.g.*Arabidopsis*), LD is predicted to extend much further than in predominantly out-crossing species such as maize (Nordborg, 2000.)

Hamblin *et al.* (2004) examined linkage disequilibrium in a diverse population of *Sorghum bicolor* for 95 loci derived from mapped RFLPs. None of the regions studied were more than 400 bp, so the decay in LD across longer regions could not be estimated. Within the 400 kb regions, however, the LD detected was approximately seven times greater than that found in maize, as described above. The authors comment that preliminary data from further studies in sorghum suggest that LD may dissipate within 10kb or less. Sorghum thus appears to be intermediate between maize and Arabidopsis in the extent of LD, consistent with its high but partial rate of self-pollination.

Recently, Caldwell *et al.* (2006) described divergent levels of LD in different barley populations. They analyzed LD within and between four genomic regions surrounding the *Ha* barley hardness locus in elite barley cultivars, a landrace population and a collection of wild ancestral material. When comparing between the four genomic regions, linkage disequilibrium extended over a region spanning 212 kb within the elite population. This reduced to 83 kb in the landrace population, and disequilibrium could not be detected in the wild population. Looking within the genomic regions showed that LD extended between 400 bp to 1100 bp in the wild material. The authors suggest that the large differences in LD between populations could be exploited in association studies by using a two-tier approach. In elite germplasm where LD extends over long regions, whole-genome scans could be used to identify candidate gene regions. High-resolution LD mapping could then be done for these regions in the landrace and wild material to identify candidate genes.

The studies described above are all concerned with measuring the physical extent of linkage disequilibrium. This provides information on the genetic architecture of populations and the effect that the measured LD will have on likelihood of detecting marker-trait associations through association analysis. Early work using an association analysis to identify trait associations was reported by Virk *et al.* (1995) in rice. A set of 47 accessions was chosen from the International Rice Research Institute (IRRI) gene bank based on phenotypic diversity for ten agronomic traits, and screened for genetic variation using seven RADP markers and 15 isoenzymes. Scored polymorphisms were used in step-wise multiple regression models with phenotypic data on culm number and days to 50% flowering as the

dependent variables. A set of 24 markers explained 99.8% of the variation in flowering time, and 13 markers explained 90% of the variation in culm number. From the models, phenotype of the two traits was predicted for each genotype and compared to the observed value. Out of the 94 trait value predictions, only four differed significantly from the observed value – one for flowering time and three for culm number. It should be noted however, that this study did not take population structure into account. As the accessions used were chosen to represent different geographic populations of rice, population stratification is highly likely. The authors note that one benefit of marker information is the efficient selection of parents for producing bi-parental mapping populations for specific QTLs of interest, in terms of having markers associated with the extremes of phenotype for the trait. This work was extended in Virk *et al.* (1996) by including the additional traits of leaf length, grain width, panicle length and culm length in the analysis.

In oats, 64 North American oat cultivars and landraces were used to study marker trait associations for 13 yield component and physiological traits using RFLP markers (Beer *et al.,* 1997). Associations were detected by pooling the accessions into groups either possessing or lacking the marker, and comparing the trait mean value for the two groups using a *t*-test. A total of 226 associations were significant at $p = 0.01$ for the traits analyzed. Thirty one markers showing significance were common to an RFLP map derived from a bi-parental cross, and could be compared. In five out of ten cases, markers associated with a high trait value in the mapping population were associated with low trait value in the germplasm pool. For the remaining QTLs, no correspondence between markers in the two populations was found. The authors conclude that the low level of congruence between QTLs identified in the two populations may limit the usefulness of the association approach in oats, and that this may be the result of genotype-by-environment interactions, multiple allele effects and marker/QTL recombination.

Kraakman *et al.* (2004) investigated associations between markers and the quantitative traits of yield and yield stability in a collection of 146 modern commercial European spring barley genotypes. As the efficiency of association-based methods depends on the extent of linkage disequilibrium, one of the objectives was to estimate the level of LD within barley. An integrated barley map derived from three segregating crosses was used to compare and validate marker-trait associations detected in the germplasm collection. The authors found that LD was common in the germplasm for markers within 10 cM distance on the integrated map. Stepwise regression selected sets of 18 to 20 markers that explained up to 58% of the variation for the traits studied. In contrast to the report by Beer *et al.* (1997), all of the yield-associated markers coincided with the chromosome bins where yield QTLs had previously

been reported (e.g. Hayes, *et al.*, 1993). This is despite the fact that most of the published QTLs were identified in North American barley, and the association study used only European germplasm. In addition, two of the identified yield stability markers coincided with a region earlier found to be involved in QTL-by-environment interactions.

Association methods using a targeted-gene approach have been used in studies of disease resistance in potato. Gebhardt *et al.* (2004) used five markers linked to known mapped QTLs for resistance to late blight and plant maturity to screen a gene bank collection of 600 potato cultivars and 114 accessions of 30 wild *Solanum* species. Two PCR markers specific for the major late blight resistance gene *R1* were highly significantly associated with resistance within the germplasm pool, as well as two anonymous PCR markers flanking the *R1* gene. Surveying the presence of the fragments in the wild *Solanum* accessions showed that the PCR amplicons associated with resistance had been introgressed into potato cultivars from the wild species *S. demissum*. The authors note that the markers can be used in breeding to screen parental clones, to select new cross combinations and to select more resistant progeny. In a similar approach, Simko *et al.* (2004b) screened 139 cultivars within the pedigree of North American potato cultivars with a microsatellite marker allele linked to a QTL for resistance to Verticillium wilt disease (VWD). Pedigree and marker analysis showed that clone USDA 41956 has at least three copies of the resistance allele, and produces a high frequency of resistant offspring. The clone USDA X96-56 lacks the allele, and frequently produces susceptible progeny. These two cultivars have made a large contribution to the VWD phenotype of North American commercial cultivars. Several tetraploid populations were developed using marker information and are currently under field trial to test the effectiveness of using the SSR allele for marker-assisted selection. In another study, Simko *et al.* (2004a) sequenced the *StVe1* gene that confers partial resistance to VWD from 30 potato cultivars in order to develop allele-specific SNP markers within the locus. Three distinct SNP haplotypes of *StVe1* were found which occurred in 97%, 33% and 10% of the germplasm studied. Although in theory heterozygous tetraploid genotypes could contain all three haplotypes, a maximum of two alleles was found in each of the North American cultivars.

Detection of QTLs for growth and forage quality traits in an admixed population of the perennial grass *Leymus* (wildryes) was reported by Hu *et al.* (2005). The admixed population was derived from F1 hybrids between two *Leymus* species, followed by two generations of open pollination. Although it is well known that admixture can cause false associations, the authors argue that in this case admixture linkage disequilibrium (ALD) should be attributable to physical linkage as false associations will be reduced by

chromosome assortment in the two generations of open pollination. Six traits were tested for association with 647 AFLP polymorphisms, and significant markers were compared with an integrated AFLP map produced from two full-sib mapping populations derived from the same initial F1 hybrids. A total of 237 markers showed positive association at p = 0.05, and half of these (119) were assigned to 37 linkage blocks spanning 13.6% of the consensus map. Twenty-eight of the strongest markers were located in only 15 linkage blocks spanning from 0.6 cM to 21.3 cM. Although in most cases linked marker associations were consistent with their predicted effect, several chromosome regions displayed apparently contradictory effects among closely linked markers. For example in some cases, the marker showed a positive effect on trait value, whereas a second or third linked marker showed a negative effect. It is possible that this is caused by ALD not attributable to physical linkage, but to genetic heterogeneity in the founder population. The authors comment that the interpretation of marker associations in this admixed population is difficult, and that additional map information from the parental population, as well as more sophisticated experimental design would be of benefit. However they also note that unless marker methods are robust, relatively simple and accessible, they are unlikely to be used by plant breeders.

Breseghello and Sorrels (2005) used a mixed-effects model for association mapping of kernel size and milling quality in wheat, in order to take population structure into account. Using prior genetic map information, population structure of 95 wheat cultivars was estimated using 36 unlinked SSR markers. The method of Pritchard *et al.* (2000a) was applied, which indicated that the population was comprised of four subpopulations. In the subsequent analysis, SSR marker was used as a fixed effect, and subpopulation as a random effect. On the three chromosomes used for association mapping, linkage disequilibrium ranged from less than 1 cM to 5 cM. A total of 14 significant associations were detected for kernel morphological traits, and six for milling quality traits. Some of the morphological associations were in regions where kernel-related QTLs had previously been identified. Kernel width was associated with marker *Xgwm30* but not with other markers closely linked to *Xgwm30*, suggesting the enhanced resolution power of association mapping compared to conventional mapping in this instance. *Xgwm30* is located on chromosome 2D, and in this study linkage disequilibrium extended on average less than 1 cM on this chromosome. The authors conclude by saying that association mapping can enhance conventional QTL studies, but that the association would need to be confirmed for individual cultivars before being used in breeding.

As second publication by Breseghello and Sorrels (2006) gives a more general account of association analysis (AA), with the objective of raising awareness among breeders of issues

related to the application of AA in breeding programmes. They define association between a QTL and a marker in terms of conditional probabilities, and compare the use of AA in three population types, *viz* germplasm collection, elite lines and synthetic populations. These population types differ for parameters such as population structure, trait heritability and expected level of linkage disequilibrium. Correspondingly, they differ in AA parameters such as statistical power of detecting association, the level of resolution possible, and the method of applying significant markers in breeding. They use the term 'association analysis' as they argue that 'association mapping' is not appropriate in the context of plant breeding populations, and reason that AA in plant breeding programmes should be considered a method of identifying markers for indirect selection, rather than a method of fine mapping QTLs.

2.7. Linkage disequilibrium and association analysis in sugarcane.

As the effective application of markers identified through association is dependent on the presence of linkage disequilibrium, the extent of LD within sugarcane breeding populations is a key issue. Although disequilibrium can be caused by a variety of other factors such as genetic drift, population structure or admixture, variable recombination and mutation rates and selection (Palmer and Cardon, 2005), disequilibrium due to physical linkage is an important population parameter influencing the outcome of association studies (Flint-Garcia *et al.*, 2003, Breseghello and Sorrels, 2006). Sugarcane breeding history is characterized by a limited number of original ancestral clones contributing the majority of germplasm currently available (Arceneaux, 1965), and a limited number of generations (~10 or less) from the original hybridizations. In fact the cultivar POJ2878 is found in the genealogy of most commercial cultivars worldwide. Due to this strong founder effect and the limited number of generations since then, it is expected that LD within sugarcane germplasm is fairly extensive. This was investigated by Jannoo *et al.* (1999) by comparing the association between RFLP markers in 59 cultivars against the RFLP map of R570. Fisher exact tests of all 2x2 marker contingency tables were performed, and 51 significant associations between pairs of markers occurring on the R570 map could be examined. In the majority of cases, LD extended over regions of 10 cM or less, although a few cases of association between markers up to 30 cM apart were found.

Identifying markers linked to phenotype of interest, and using these markers as tools in breeding and selection, has been the goal behind much of the work in sugarcane molecular analysis to date, as far back as the work of Roughan *et al.* (1971), reviewed above. The next attempt to identify markers for a specific trait in sugarcane – stalk fibre content - was

reported by Msomi and Botha (1994). This preliminary report describes the use of bulk segregant analysis (BSA) methodology in a segregating population of a bi-parental cross to identify candidate RAPD markers for further screening and validation across individual progeny. Although the population used were progeny from a bi-parental cross, this study is also the first report on marker identification that does not rely directly on a mapping approach in sugarcane. Msomi (1998) later reported identifying three markers ascribing 4.8%, 9.2% and 14.6% of the phenotypic variation for fibre content in a progeny population of 80 individuals. An attempt was made to convert these to markers based on sequence characterised amplified regions (SCARs), but these proved to be monomorphic across the study population.

A second attempt to identify markers for traits through statistical association rather than by mapping was reported by Barnes *et al.* (1997). In this study, a population of 50 sugarcane genotypes with known phenotypic ratings for resistance to smut, eldana and sugarcane mosaic virus (SCMV) were screened for DNA polymorphisms with 41 random primers. Stepwise regression using the resulting 382 scored RAPD markers accounted for 34.7%, 40.1% and 31.6% of the phenotypic variation for eldana, SCMV and smut, with models consisting of four, four and three markers respectively. Although this study was based on a fairly small number of genotypes, and the fact that the RAPD technique has since fallen out of favour due to difficulties in the reproducibility of results, this initial study did illustrate that marker identification through association was possible in sugarcane, and worthy of further consideration. Part of the work described in the following chapters is based on this approach, using a larger population of genotypes and different marker systems.

McIntyre *et al.* (2005) used an association approach to validate markers identified through a mapping study in sugarcane. In a mapping population of progeny of two elite sugarcane clones (Aitken *et al.*, 2005), resistance to pachymetra root rot and brown rust was measured over two years. 13 markers were associated with pachymetra root rot in at least one year, and 15 markers were associated with rust. To determine whether they would be useful in a broader genetic background, these markers were screened across a set of 154 elite sugarcane clones. Six of the 13 pachymetra markers remained associated, and seven of the 15 rust markers. The results suggested that these markers could be useful for selection among the broader sugarcane population.

From the studies above, it is evident that association or linkage disequilibrium methods are becoming more commonly tested in different crop species. A common element between all the reviewed applications, however, is that they have relied on existing map information to

target genomic areas for more detailed molecular dissection, or to interpret the molecular data. The sugarcane work reported in the rest of this thesis will be following a somewhat different two-phase approach. The first objective is to exploit the fairly long-range LD detected in sugarcane (Jannoo *et al.*, 1999) to do a low resolution genome scan for regions of interest associated with pest and disease resistance loci, similar to the strategy suggested by Caldwell *et al.* (2006).

The second phase will be to use the association between markers to construct a population-level map of genetic regions in disequilibrium, that co-segregate within the population due to physical linkage or other causes of association. Creating a map *de novo* using association data is a novel approach not reported to date for sugarcane or any other crop species. Although this low resolution map will not be appropriate for fine mapping or map-based cloning of genes of interest, it is believed that the information will be useful as a practical tool in guiding breeding and selection decisions in an applied variety improvement programme.

2.8. References.

Aitken, KS, Jackson, PA and McIntyre, CL. 2005. A combination of AFLP and SSR markers provides extensive map coverage and identification of homo(eo)logous linkage groups in a sugarcane cultivar. Theoretical and Applied Genetics 110: 789-801.

Aitken, KS, Jackson, PA and McIntyre, CL. 2006. QTL identified for sugar related traits in a sugarcane (*Saccharum* spp.) cultivar x *S. officinarum* population. Theoretical and Applied Genetics 112: 1306-1317.

Al-Janabi, SM, Honeycutt, RJ, McClelland, M and Sobral, BWS. 1993. A genetic linkage map of *Saccharum spontaneum* L. 'SES 208'. Genetics 134: 1249-1260.

Al-Janabi, S.M, Honeycutt, RJ, and Sobral, BWS. 1994. Chromosome assortment in *Saccharum*. Theoretical and Applied Genetics 89: 959-963.

Arceneaux, G. 1965. Cultivated sugarcanes of the world and their botanical derivation. Proceedings of the International Society of Sugar Cane Technologists. 12: 844-854.

Asnaghi, C, Roques, D, Ruffel, S, Kaye, C, Hourau, JY, Telismart, H, Girard, JC, Raboin, LM, Risterucci, AM, Grivet, L and D'Hont, A. 2004. Targetted mapping of a sugarcane rust resistance gene (*Bru1*) using bulked segregant analysis and AFLP markers. Theoretical and Applied Genetics 108: 759-764.

Balding, DJ, Bishop, M and Cannings, C (eds). 2003. Handbook of Statistical Genetics, 2nd edition. Volume 1. John Wiley and Sons, Ltd. pp 574.

Barnes, JM, Rutherford, RS and Botha, FC. 1997. The identification of potential genetic markers in sugarcane varieties for the prediction of disease and pest resistance ratings. Proceedings of the South African Sugar Technologists Association 71: 57-61.

Beer, C, Siripoonwiwat, W, O'Donoughue, LS, Souza, E, Matthews, D and M.E. Sorrells, ME. 1997. Associations between molecular markers and quantitative traits in an oat germplasm pool: Can we infer linkages? Journal of Agricultural Genomics. Published with permission from CAB International. Full text available from http://www.cabi-publishing.org/jag/papers97/paper197/indexp197.html

Berding, N and Roach, BT. 1987. Germplasm collection, maintenance and use. In: Sugarcane Improvement through Breeding, Ed: Heinze, DJ. Elsevier, Amsterdam. pp 143-210.

Breseghello, F and Sorrels, ME. 2005. Association mapping of kernel size and milling quality in wheat. Genetics 172: 1165-1177.

Breseghello, F and Sorrels, ME. 2006. Association analysis as a strategy for improvement of quantitative traits in plants. Crop Science 46: 1323-1330.

Caldwell, KS, Russell, J, Langridge, P and Powell, W. 2006. Extreme population-dependent disequilibrium detected in an inbreeding plant species, *Hordeum vulgare*. Genetics 172: 557-567.

Cardon, LR and Palmer, L. 2003. Population stratification and spurious allelic association. The Lancet 361: 598-604.

Ching, A, Caldwell, KS, Jung, M, Dolan, M, Smith, OS, Tingey, S, Morgante, M and Rafalski, AJ. 2002. SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BioMed Central Genetics 3: 19-32.

Collins, A and Morton, NE. 1998. Mapping a disease locus by allelic association. Proceedings of the National Academy of Sciences USA 95: 1741-1745.

Daugrois, JH, Grivet, L, Roques, D, Hoarau, JY, Lombard, H, Glaszmann, JC and D'Hont, A. 1996. A putative major gene for rust resistance linked with a RFLP marker in sugarcane cultivar 'R570'. Theoretical and Applied Genetics 92: 1059-1064.

Da Silva, J, Sorrells, ME, Burnquist, W and Tanksley, SD. 1993. RFLP linkage map and genome analysis of *Saccharum spontaneum*. Genome 36: 782-791.

Da Silva, J, Honeycutt, RJ, Burnquist, W, Al-Janabi, SM, Sorrells, ME, Tanksley, SD and Sobral, BWS. 1995. *Saccharum spontaneum* L. "SES208' genetic linkage map combining RFLP and PCR-based markers. Molecular Breeding 1: 165-179.

D'Hont, A, Lu, Y, Gonzales de Leon, D, Grivet, L, Feldmann, P, Lenaud, C and Glaszmann, JC. 1994. A molecular approach to unraveling the genetics of sugarcane, a complex polyploid of the Andropogoneae tribe. Genome 37: 222-230.

D'Hont, A, Grivet, L, Feldmann, P, Rao, S, Berding, N and Glaszmann, JC. 1996. Characterisation of the double genome structure of modern sugarcane cultivars (*Saccharum* spp.) by molecular cytogenetics. Molecular and General Genetics 250: 405-413.

D'Hont, A, Ison, D, Alix, K, Roux, C and Glaszmann, JC. 1998. Determination of the basic chromosome numbers in the genus Saccharum by physical mapping of ribosomal RNA genes. Genome 41: 221-225.

Ewens, WJ and Spielman, RS. 1995. The transmission/disequilibrium test: history, subdivision and admixture. American Journal of Human Genetics 57: 455-464.

Flint-Garcia, SA, Thornsberry, J and Buckler, ES. 2003. Structure of linkage disequilibrium in plants. Annual review of plant biology 54: 357-374.

Gebhardt, C, Ballvora, A, Walkemeier, B, Oberhagemann, P and Schuller, K. 2004. Assessing genetic potential in germplasm collections of crop plants by marker-trait association: a case study for potatoes with quantitative variation of resistance to late blight and maturity type. Molecular Breeding 13: 93-102.

Glaszmann, JC, Fautret, A., Noyer, JL., Feldmann, P and Lanaud, C. 1989. Biochemical genetic markers in sugarcane. Theoretical and Applied Genetics 78: 537-543.

Goldstein, DB and Weale, ME. 2001. Population genomics: Linkage disequilibrium holds the key. Current Biology 11: R576-R579.

Grivet, L, D'Hont, A, Dufour, P, Hamon, P, Roques, D and Glaszmann, JC. 1994. Comparative genome mapping of sugarcane with other species within the Andropogoneae tribe. Heredity 73: 500-508.

Grivet, L, D'Hont, A, Roques, D, Feldmann, P, Lanaud, C and Glaszmann, JC. 1996. RFLP mapping in cultivated sugarcane (*Saccharum* spp): Genome organization in a highly polyploid and aneuploid interspecific hybrid. Genetics 142: 987-1000.

Guimaraes, CT, Sills, GR and Sobral, BWS. 1997. Comparative mapping of Andropogoneae: *Saccharum* L. (sugarcane) and its relation to sorghum and maize. Proceedings of the National Academy of Sciences USA 94: 14262-14266.

Hagenblad, J and Nordborg, M. 2002. Sequence variation and haplotype structure surrounding the flowering time locus *FRI* in *Arabidopsis thaliana*. Genetics 161: 289-298.

Hamblin, MT, Mitchell, SE, White, GM, Gallego, J, Kukatla, R, Wing, RA, Paterson, AH and Kresovich, S. 2004. Comparative population genetics of the Panicoid grasses: sequence polymorphism, linkage disequilibrium and selection in a diverse sample of *Sorghum bicolor*. Genetics 167: 471-483.

Hayes, PM, Liu, BH, Knapp, SJ, Chen, F, Jones, B, Blake, T, Franckowiak, J, Rasmusson, D, Sorrels, M, Ullrich, SE, Wesenberg, D and Kleinhofs, A. 1993. Quantitative trait locus effects and environmental interaction in a sample for North American barley germplasm. Theoretical and Applied Genetics 87: 392-401.

Heinz, DJ. 1969. Isoenzyme prints for variety identification. ISSCT sugarcane breeders newsletters 24:8.

Hoarau, JY, Offmann, B, D'Hont, A, Risterucci, AM, Roques, D, Glaszmann, JC and Grivet, L. 2001. Genetic dissection of a modern cultivar (*Saccharum* spp.). I. Genome mapping with AFLP. Theoretical and Applied Genetics 103: 84-97.

Hoarau, JY, Grivet, L, Offmann, B, Raboin, LM, Diorflat, JP, Payet, J, Hellmann, M, D'Hont, A and Glaszmann, JC. 2002. Genetic dissection of a modern cultivar (*Saccharum* spp.). II. Detection of QTLs for yield components. Theoretical and Applied Genetics 105: 1027-1037.

Hu, ZM, Wu, XL, Larson, SR, Wang, RRC, Jones, TA, Chatterton, NJ and Palazzo, AJ. 2005. Detection of linkage disequilibrium QTLs controlling low-temperature growth and metabolite accumulations in an admixed breeding population of *Leymus* wildryes. Euphytica 141: 263-280.

Jannink, JL, Bink, MCAM, and Jansen, RC. 2001. Using complex plant pedigrees to map valuable genes. Trends in Plant Science 6: 337-342.

Jannoo, N, Grivet, L, Dookun, A, D'Hont, A and Glaszmann, JC. 1999. Linkage disequilibrium among sugarcane cultivars. Theoretical and Applied Genetics 99: 1053-1060.

Jordan, DR, Casu, RE, Besse, P, Carroll, BC, Berding, N and McIntyre, CL. 2004. Markers associated with stalk number and suckering in sugarcane co-locate with tillering and rhizomatousness QTLs in sorghum. Genome 47: 988-993.

Kraakman, ATW, Niks, RE, Van den Berg, PMMM, Stam, P and Van Eeuwijk, FA. 2004. Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars. Genetics 168: 435-446.

Lander, ES and Schork, NJ. 1994. Genetic dissection of complex traits. Science 265: 2037-2048.

Lindstrom, EW. 1924. A genetic linkage between size and colour factors in tomato. Science 60: 182-183.

Lindstrom, EW. 1931. Genetic tests for linkage between row number and certain qualitative genes in maize. Research Bulletin of Iowa State College of Agriculture 142: 250-288.

Liu, BH. 1998. Statistical Genomics: Linkage, mapping and QTL analysis. CRC Press, LLC. pp 611.

Lynch, M and Walsh, B. 1998. Genetics and Analysis of Quantitative Traits. Sinauer Associates Inc. Sunderland, Massachusetts. pp 980.

McIntyre, CL, Whan, VA, Croft, B, Magarey, R and Smith, GR. 2005. Identification and validation of molecular markers associated with Pachymetra root rot and brown rust in sugarcane using map and association based approaches. Molecular Breeding 16: 151-161.

Ming, R, Liu, SC, Lin, YR, da Silva, J, Wilson, W, Braga, D, van Deynze, A, Wenslaff, TF, Wu, KK, Moore, PH, Burnquist, W, Sorrells, ME, Irvine, JE and Paterson, AH. 1998. Detailed alignment of *Saccharum* and *Sorghum* chromosomes: Comparative organization of closely related diploid and polyploid genomes. Genetics 150: 1663-1682.

Ming, R, Liu, SC, Moore, PH, Irvine, JE and Paterson, AH. 2001. QTL analysis in a complex autopolyploid: Genetic control of sugar content in sugarcane. Genome Research 11: 2075-2084.

Ming, R, Liu, SC, Bowers, JE, Moore, PH, Irvine, JE and Paterson, AH. 2002a. Construction of a *Saccharum* consensus map from two interspecific crosses. Crop Science 42: 570-583.

Ming, R, Wang, YW, Draye, X, Moore, PH, Irvine, JE and Paterson, AH. 2002b. Molecular dissection of complex traits in autoployploids: mapping QTLs affecting sugar yield and related traits in sugarcane. Theoretical and Applied Genetics: 332-345.

Ming, R, Del Monte, TA, Hernandez, E, Moore, PH, Irvine, JE, and Paterson, AH. 2002c. Comparative analysis of QTLs affecting plant height and flowering among closely-related diploid and polyploid genomes. Genome 45: 794-803.

Morton, NE, Zhang, W, Taillon-Miller, P, Ennis, S, Kwok, PY and Collins, A. 2001. The optimal measure of allelic association. Proceedings of the National Academy of Sciences USA 98: 5217-5221.

Msomi, N and Botha, FC. 1994. Identification of molecular markers linked to fibre using bulk segregant analysis. Proceedings of the South African Sugar Technologists Association 68:41-45.

Msomi, NS. 1998. The potential of bulk segregant analysis and RADP technology for the identification of molecular markers liked to traits in sugarcane. Unpublished Ph.D thesis, University of Natal, Durban, South Africa.

Mudge, J, Anderson, WR, Kehrer, RL and Fairbanks, DJ. 1996. A RAPD genetic map of *Saccharum officinarum*. Crop Science 36: 1362-1366.

Nordborg, M. 2000. Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. Genetics 154. 923-929.

Nordborg, M, Borevitz, JO, Bergelson, J, Berry, CC, Chory, J, Hagenblad, J, Kreitman, M, Maloof, JN, Noyes, T, Oefner, PJ, Stahl, EA and Weigel, D. 2002. The extent of linkage disequilibrium in *Arabidopsis thaliana*. Nature Genetics 30:190-193.

Palmer, LJ and Cardon, LR. 2005. Shaking the tree: mapping complex disease genes with linkage disequilibrium. The Lancet 366: 1223-1234.

Payne, F. 1918. The effect of artificial selection on bristle number in *Drosophila ampelophila* and its interpretation. Proceedings of the National Academy of Sciences USA 4: 55-58.

Price, S. 1967. Interspecific hybridization in sugarcane breeding. Proceedings of the International Society of Sugar Cane Technologists 12: 1021-1026.

Pritchard, JK and Przeworski, M. 2001. Linkage disequilibrium in Humans: Models and data. American Journal of Human Genetics 69: 1-14.

Pritchard, JK and Rosenberg, NA. 1999. Use of unlinked genetic markers to detect population stratification in association studies. American Journal of Human Genetics 65: 220-228.

Pritchard, JK, Stephens, M and Donnelly, P. 2000a. Inference of population structure using multilocus genotype data. Genetics 155: 945-959.

Pritchard, JK, Stephens, M, Rosenberg, NA and Donnelly, P. 2000b. Association mapping in structured populations. American Journal of Human Genetics 67: 170-181.

Pritchard, JK and Wen, W. 2003. Documentation for structure software: Version 2. http://pritch.bsd.uchicago.edu/software.html. Verified 1 November 2005.

Rafalski, A. 2002. Applications of single nucleotide polymorphisms in crop genetics. Current Opinion in Plant Biology 5: 94-100.

Rasmusson, J. 1927. Genetically changed linkage values in *Pisum*. Hereditas 10: 1-152.

Reffay, N, Jackson, PA, Aitken, KS, Hoarau, JY, D'Hont, A, Besse, P and McIntyre, CL. 2005. Characterisation of genome regions incorporated from an important wild relative into Australian sugarcane. Molecular Breeding 15: 367-381.

Reich, DE, Cargill, M, Bolk, S, Ireland, J, Sabeti, PC, Richter, DJ, Lavery, T, Kouyoumjian, R, Farhadian, SF, Ward, R and Lander, ES. 2001. Linkage disequilibrium in the human genome. Nature 411: 199-204.

Remington, DL, Thornsberry, JM, Matsuoka, Y, Wilson, LM, Whitt, SR, Doebley, J, Kresovich, S, Goodman, MM and Buckler, ES. 2001. Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proceedings of the National Academy of Sciences USA 98: 11479-11484.

Risch, N and Merikangas, K. 1996. The future of genetic studies of complex human diseases. Science 273: 1516-1517.

Roughan, PG., Waldron, JC and Glasziou, KT. 1971. Starch inheritance in *Saccharum*. Enzyme polymorphism for β–amylase in interspecific and intergeneric hybrids. Proceedings of the International Society of Sugar Cane Technologists 14: 257-265.

Roach, BT and Daniels, J. 1987. A review of the origin and improvement of sugarcane. In: Copersucar International Sugarcane Breed Workshop. Sao Paulo. pp1-32.

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

Simko, I, Haynes, KG, Ewing, EE, Costanzo, S, Christ, BJ and Jones, RW. 2004a. Mapping genes for resistance to *Verticillium albo-atrum* in tetraploid and diploid potato populations using haplotype association tests and genetic linkage analysis. Molecular Genetics and Genomics 271: 522-531.

Simko, I, Haynes, KG and Jones, RW. 2004b. Mining data from potato pedigrees: tracking the origin of susceptibility and resistance to *Verticillium dahliae* in North American cultivars through molecular marker analysis. Theoretical and Applied Genetics 108: 225-230.

Simmonds, NW. 1976. Sugarcanes. In: Evolution of Crop Plants. Ed: Simmonds, NW. Longmans, London. pp 104-108.

Spielman, R, McGinnis, R and Ewens, W. 1993. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). American Journal of Human Genetics 52: 506-516.

Stevenson, GC. 1965. Genetics and Breeding of Sugarcane. Tropical Science Series. Longmans, Green and Co Ltd. London. 284pp.

Tanksley, SD, Grandillo, S, Fulton, TM, Zamir, D, Eshed, Y, Petiard, V, Lopez, J and Beck-Bunn, T. 1996. Advanced backcross QTL analysis in a cross between an elite processing line of tomato and its relative *L. pimpinellifolium*. Theoretical and Applied Genetics 92: 213-224.

Tenaillon, MI, Sawkins, MC, Long, AD, Gaut, RL, Doebley, JF and Gaut, BS. 2001. Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.) Proceedings of the National Academy of Sciences USA 98: 9161-9166.

Thornsberry, JM, Goodman, MM, Doebley, J, Kresovich, S, Nielson, D and Buckler, ES. 2001. *Dwarf8* polymorphisms associated with variation in flowering time. Nature Genetics 28: 286-289.

Virk, PS, Ford-Lloyd, BV, Jackson, MT, Pooni, HS, Clemeno, TP and Newbury, HJ. 1995. Marker-assisted prediction of agronomic traits using diverse rice germplasm. Rice genetics III. Proceedings of the Third International Rice Genetics Symposium, 16-20 Oct 1995. Manila (Philippines): pp 307-316. Available from www.irri.org/science/abstracts/pdfs/RGIIIMarker26.pdf. Verified 1 Nov 2005.

Virk, PS, Ford-Lloyd, BV, Jackson, MT and Newbury, HJ. 1996. Predicting quantitative variation within rice germplasm using molecular markers. Heredity 76: 296-304.

Wexelsen, H. 1934. Linkage between quantitative and qualitative traits in barley. Hereditas 17: 323-341.

Wu, KK, Burnquist, W, Sorrells, ME, Tew, TL, Moore, PH and Tanksley, SD. 1992. The detection and estimation of linkage in polyploids using single-dose restriction fragments. Theoretical and Applied Genetics 83: 294-300.

CHAPTER 3

Identification of molecular markers associated with response to infection by smut and attack by eldana within a sugarcane breeding population

3.1. Introduction.

In South Africa, the African sugarcane stalk borer *Eldana saccharina* Walker (Lepidoptera: Pyralidae) is the major pest of sugarcane, causing extensive damage especially when cane is exposed to abiotic stress such as drought. As the larvae that cause the damage are protected within the stalk, chemical control is largely ineffective, and host-plant resistance has been the only method available to limit the level of damage. Although stalk fibre content is an effective resistance mechanism against eldana, it is an undesirable trait in terms of the milling process. Non-mechanical forms of resistance mechanisms such as antixenosis and antibiosis are therefore more desirable traits from a commercial perspective (Keeping and Rutherford, 2004). These traits are poorly understood, however, and difficult to characterize, and selection for resistance relies on field and greenhouse screening of borer damage in large numbers of genotypes. High levels of resistance not associated with high stalk fibre content are rare within the breeding population, and this remains one of the major challenges to the SASRI breeding programme.

The smut fungus (*Ustilago scitaminea,* Syd.) is one of the important diseases of sugarcane that limits the release of high yielding varieties (Butterfield and Thomas, 1996). Within the breeding population, resistance to smut and eldana appear to be negatively correlated (SASRI, unpublished data). Both smut and eldana often enter the stalk through the bud, and it has been hypothesized that separate but interacting plant defense responses against insects and microbial pathogens may be implicated. Studies in other plants have suggested that negative interaction between the jasmonate and salicylate signaling pathways may be responsible for the contrasting reaction of plants to attack by herbivores or pathogens (Thaler *et al.*, 2002, Spoel *et al.*, 2003). Whatever the actual mechanism, breeding for resistance to both smut and eldana are important goals of the SASRI breeding program; a task made more difficult by the negative correlation between the traits. Because of this, molecular markers linked to smut and/or eldana resistance could be valuable tools to improve the efficiency of breeding commercially desirable cultivars.

The SASRI sugarcane breeding programme is carried out at five different sites selected to represent the major climatic and soil environments represented within the South African sugar industry. Crossing operations are centralized at the Sugarcane Research Institute in Mt Edgecombe, and each year approximately 1500 crosses are made between around 250 different parent genotypes. The genotypes used as parents are selected based on their adaptation to the regions represented by the five selection sites in terms of phenotypic performance for yield, sucrose and pest and disease traits. Sugarcane flowering physiology is affected by both photoperiod and temperature (Moore and Nuss, 1987) and flowers used in crossing have to be produced under controlled conditions in photoperiod facilities and a greenhouse in order to ensure that they are fertile. Space limitations in the facilities mean that the number of parent genotypes that can be used each year is limited, and desirable cross combinations need to be identified in advance in order to choose which parents to plant inside the photoperiod and glasshouse facilities. Following crossing, approximately 50 000 potted seedlings from 100-150 different families (crosses) are sown each year at each of the five selection sites. These seedlings then undergo a five-stage screening programme, as outlined in Table 3.1.

Table 3.1. General outline of the variety selection programme at each SASRI selection site.

Selection stage	No. genotypes	No. plots and replications	No. crop cycles
Potted seedlings	50000	single plant	1
Stage 1	35000	single plant	1
Stage 2	4000	1row x 8m	2
Stage 3	400	2rows x 8m x 2 reps	1
Stage 4	60	5rows x 8m x 3 reps	3
Stage 5	25	5rows x 8m x 3 reps x 4 locations	3

Due to the large numbers of genotypes, selection at Stage 1 is based on a visual assessment of vigour, as well as lack of visual symptoms of pests and diseases. From Stage 2 onwards all plots are harvested, weighed and sampled for sucrose and cane quality traits in each crop cycle, and estimation of yield potential improves as the plot size and number of replications increases. Surveys for pests and diseases are carried out in all selection stages, and genotypes in Stages 3, 4 and 5 are sent to separate trials conducted

by the Pathology and Entomology sections to assess resistance to eldana, smut and sugarcane mosaic virus. Phenotypic ratings of resistance/susceptibility are assigned based on the scale published by the International Society of Sugar Cane Technologists (ISSCT) (Hutchinson, 1968), with a score of 1 equivalent to highly resistant, 5 being intermediate and 9 being highly susceptible. The length of time taken to complete the selection process is dependent on the length of the cropping cycle at each location, and it takes between 11 to 15 years before a candidate genotype is released as a commercial variety. All yield, cane quality and pest and disease data derived during the testing period is stored in an Oracle database for easy retrieval.

Although varieties highly susceptible to smut and/or eldana may be identified early in the selection programme and discarded, reliable information on resistance is only available towards the end of the selection programme, once genotypes have been through several screening trials for smut and eldana. Having reliable molecular markers linked to genes controlling resistance to smut and/ or eldana would be of great benefit in improving the efficiency of combined breeding for yield improvement and pest/disease resistance. However the large numbers of genotypes in the seedling and early selection stages makes screening these populations for marker-assisted selection (MAS) impractical from a logistical and cost perspective. As an alternative strategy to conventional MAS, molecular marker information could be used to pyramid different resistance genes or alleles into progeny populations by selecting specific cross combinations based on the molecular marker profiles of the parent genotypes. This would be possible if marker information was available for a large number of parent genotypes used in crossing. In order to be effective, a gene and/or allele pyramiding (GAP) strategy requires a different marker identification strategy than the conventional approach of analyzing segregating progeny from a single cross.

Sugarcane is a highly heterozygous polyploid (Butterfield *et al.*, 2001), and it is likely that allelic variation at quantitative trait loci (QTL) is high. Several or many different quantitative trait alleles (QTAs) may be present within a single individual for those QTLs involved in the trait of interest. Conventional marker identification strategies through analyzing segregating progeny from a bi-parental cross will only be capable of identifying those QTAs present in two parent genotypes of the segregating population, and not detect the allelic variation present for exploitation within the breeding population as a whole. In addition, marker and map information for a small number of genotypes will not be particularly useful in a GAP breeding strategy, where information is required on large numbers of parents. The approach of using marker-trait association within a diverse population of germplasm to identify QTLs influencing phenotype has been used previously in other crops such as oats (Beer *et al.*,

1997), rice (Virk *et al.*, 1996), maize (Thornsberry *et al.*, 2001) barley (Igartua *et al.*, 1999, Kraakman *et al.*, 2004) and potato (Simko *et al.*, 2004). Using a similar approach in sugarcane will facilitate the identification of a broad range of allelic variants or QTAs associated with resistance to smut and eldana. In addition, if the population used for marker discovery is comprised of genotypes making up the parental germplasm pool, the resulting information can be directly applied in a GAP breeding strategy, as marker information will be available for the parents in the marker discovery population. A third benefit of this approach is that in the SASRI breeding programme, parents are selected from the large pool of genotypes that have been through the selection programme, and already have a wealth of reliable phenotypic measurements available that have been collected over different locations and different years. There is therefore no need to establish specific trials to measure the trait of interest, and the effects of genotype by environment interactions have been 'smoothed' over locations and years.

The objective of this chapter is to assess the potential and efficacy of using marker-trait association to identify molecular markers for smut and eldana resistance in a population of genotypes used as parents in the SASRI breeding programme. Previous QTA identification studies in sugarcane have used bi-parental or selfed progeny populations to identify markers for a variety of traits and these studies will be discussed in relation to the results obtained.

3.2. Materials and methods.

3.2.1. Population composition.

The population used for marker identification was 78 genotypes that had or were being used as parents in the breeding program. One genotype was later shown to be misidentified (see 3.3.1. below), and was dropped from the list. The list of 77 genotypes remaining is shown in Table 3.2. The 'NCo' varieties were the first sugarcanes released from the SASRI programme highly adapted to South African conditions, and form the foundation of the SASRI sugarcane germplasm, featuring prominently in the genealogies of later generations. Genotypes with a name beginning with a two-digit number followed by a letter are derived from the SASRI breeding programme but have not been released as commercial varieties. The two-digit number reflects the year they entered stage 2 of the selection programme, and the letter reflects the selection site code - .e.g. 68W1049 was the 1049th genotype planted in stage 2 in 1968 at site W (Shaka's Kraal on the KZN North Coast). Although not released commercially, these genotypes have been used as parents in the breeding programme as they have some desirable phenotypic traits. The 'N' varieties – viz N8 to N34 are varieties that have been commercially released from SASRI, from 1973 (N8) to 2001 (N34). The

remaining genotypes are older foreign varieties that have been used as parents in the SASRI programme – among many other foreign varieties – in order to introduce additional genetic variation into the breeding programme. Their origin is shown in Table 3.2. One of the main reasons for choosing this set of genotypes was that some molecular marker data was already available from a project initiated within the Biotechnology section at SASRI (see section 3.2.2 below).

Table 3.2. Genotypes comprising the marker identification population, with their phenotypic ratings for smut (S) and eldana (E)

Genotype	S	E	Genotype	S	E	Genotype	S	E	Genotype	S	E
NCo293	9	8	78F0909	2	7	85F1628	3	9	N22	4	4
NCo339	3	5	79F1043	8	6	85F2805	8	7	N23	3	6
NCo376	8	5	79F1855	8	3	85H0605	8	4	N24	4	6
N52/219	3	8	79H0181	5	8	85L1041	7	5	N25	6	6
NM214	3	5	79L0181	8	5	85L1056	4	5	N26	3	7
68W1049	6	6	79L1294	8	4	85L1769	6	4	N27	3	6
73L1295	8	5	79M0955	8	5	85W1610	7	4	N30	4	7
74M659	5	4	80E1496	7	4	87L0329	3	6	N31	7	6
75E0247	4	4	80F2147	6	8	87L1484	5	3	N32	6	5
75E1293	6	6	80L0432	3	7	88W1323	8	8	N33	5	3
75L1157	8	3	80L0627	5	7	N8	9	3	N34	6	5
75L1463	8	3	80M1257	7	2	N11	3	8	**Foreign varieties**		
76H0333	3	8	80W1459	8	3	N13	8	4	B42231	5	3
76M1101	5	3	81L1308	5	7	N14	5	5	CB38/22	3	3
76M1566	6	3	81W0133	8	5	N16	8	5	CB40/35	3	9
77F0637	8	4	81W0447	6	4	N17	5	3	CO281	7	3
77F0790	6	6	82F0675	3	6	N18	6	7	CO285	5	3
77L1143	8	8	82F2907	4	6	N19	5	4	CP57/614	5	3
77L1720	7	4	83F0448	3	7	N20	8	3	J59/3	3	9
77W1241	8	5	84F2753	3	8	N21	4	3			

Phenotypic data on resistance rating to smut and eldana had been collected on these genotypes over several sites and years, and was extracted from the SASRI database. For the SASRI varieties, data were derived from at least 15 trial x crop cycles, as well as two eldana inoculation trials and two smut inoculation trials conducted in different years. For the foreign varieties the data resource is more limited, and restricted to smut and eldana surveys conducted over two years in open quarantine, and three trial x crop cycle series. Ratings were assigned using the International Society of Sugar Cane Technologist (ISSCT) scale of 1 to 9 (Hutchinson, 1968), with 1 being highly resistant, and 9 being highly susceptible, and are shown in Table 3.2. DNA was extracted from fresh leaf-roll of all genotypes using established protocols (Dellaporta *et al.*, 1983).

3.2.2. RFLP markers.

As mentioned in 3.2.1 (above), part of the justification for using the populations shown in Table 3.2 was that some molecular characterization had already been done on these genotypes as part of a broader SASRI project investigating expressed sequence tags (ESTs) to identify genes involved in important plant processes. This work has been described previously (Heinze *et al.,* 2001, Thokoane and Rutherford, 2001, Butterfield *et al.*, 2004). To summarize briefly, ESTs were identified through differential display cDNA-AFLP and suppression subtractive hybridization (SSH) performed on resistant and susceptible sugarcane varieties that were either inoculated with the smut fungus, or uninfected. ESTs differentially expressed between resistant/susceptible genotypes and/or inoculated/uninfected were cloned and sequenced, and BLAST searched to identify fragments with homology to known genes involved in plant defense responses. Forty-five ESTs were selected and used as restriction fragment length polymorphism (RFLP) probes in Southern analysis of the 78 genotypes using two restriction enzymes; HindIII and Dra1. Because of the ploidy level of sugarcane, markers were scored in a dominant manner as present (1) or absent (0). A list of the probes and their putative homology is given in Table 3.3.

3.2.3. AFLP markers.

Because of the high ploidy level of sugarcane and the large genome size (~112 chromosomes, >17000cM, Hoarau *et al.*, 2001), a large number of molecular markers are required to give an adequate coverage of the genome. The cost and time required to generate large numbers of RFLP markers is prohibitive. To complement the existing RFLP data, the high-throughput Amplified Fragment Length Polymorphism (AFLP, Vos *et al.*, 1995) technique was used to rapidly generate larger numbers of markers to identify potential marker-trait associations. The AFLP work was conducted in the laboratories of CIRAD (Centre de Coopération Internationale en Recherche Agronomique pour le Développement) in Montpellier, France, and funded through a grant from the South African National Research Foundation (GUN 2065288), under the auspices of the France-South Africa Science and Technology Agreement. A set of 64 AFLP primer combinations was used to marker-type the 78 genotypes using standard protocols (Vos *et al.*, 1995) with slight modification as suggested by Hoarau *et al.* (2001) and the manufacturers instructions for γP^{33} labeling using the Gibco BRL kit. The individual AFLP primers used are shown in Table 3.4.

Table 3.3. RFLP probes used for marker generation, with their corresponding putative homology.

RFLP probe code	Putative homology
ABC1	ATP binding cassette transporter
ABC2	ATP binding cassette transporter
AscOxi	Ascorbate oxidase
AspTra	Aspartate carbamoyltransferase
BetDeh	Betain aldehyde dehydrogenase
BGlu	B-Glucosidase
CalRet	Calreticulin precursor
CelSyn	Cellulose synthase
ChalRed	Chalcone reductase
ChlCha	Chloroplast chaperone
FatRed	Fatty acid reductase
GPrRec	G protein receptor
IsoRed	Isoflavone reductase
Jac	Jacalin
LipTra	Lipid Transport
MAD522	MADS-box transcription factor
NadRed	NADH Oxido-Reductase
NbsNo1	NBS-LRR no.1
NbsNo2	NBS-LRR no.2
NbsSus	NBS-LRR sus7
OliTra	Oligosaccharide transferase
P1	Transcription factor in flavonoid pathway, from maize.
PatEst	Pathogen induced EST
Perox	Peroxidase
Pho310	Phosphoprotein phosphatase-2
PhoPho	Phosphoprotein phosphatase-1
PRZnFi	PR ZN finger
PrPro	Protein secretion
Pro20	20S Proteosome beta subunit
PtoKin	Serine/threonine protein kinase (Pto-like)
R1	Transcription factor in flavonoid pathway, from maize.
RecKin	Receptor kinase
RinZn	Ring zinc finger protein
SerKin	Serine/threonine kinase
SerInh	Serine protease inhibitor
Tha	Thaumatin
Tomyb1	Target of myb1-like protein
TMPro	Trans membrane protein
Umc106	UMC 106 - from University of Missouri
Unk301	Unknown, derived from variety Co301
Unk525	Unknown derived from variety N52/219
Unk53	Unknown derived from cDNA-AFLP
Ves	Vesicle associated membrane protein
WAK	Wall associated kinase
X1	Transcription factor in flavonoid pathway, from sugarcane

Table 3.4. AFLP primers used for marker generation. All 8x8 = 64 combinations were used.

***Eco*R1-primer**		***Mse*1-primer**	
Code	Selective nucleotide	Code	Selective nucleotide
1	AAC	1	CAA
2	AAG	2	CAC
3	ACA	3	CAG
4	ACC	4	CAT
5	ACG	5	CTA
6	ACT	6	CTC
7	AGC	7	CTG
8	AGG	8	CTT

Restriction and ligation of AFLP adaptors was performed using 250ng of DNA per genotype using the Invitrogen kit 10482-016. Pre-amplification was done on a 1/10 dilution of the restriction/ligation mix using Invitrogen kit 10792-018. Pre-amplification products were diluted 1/20 with water for the final amplification with appropriate *Eco*R1 and *Mse*1 primers. Four microlitres of reaction products were loaded onto 5% denaturing polyacrylamide gels and electrophoresis was done at 70W for one-and-a-half to two hours. Gels were transferred onto filter paper and dried at 80°C before being exposed to X-ray film for five to ten days. Two AFLP gels for each primer combination were required to analyze all the genotypes in the population, resulting in a total of 128 AFLP autoradiographs. In order to try and facilitate fragment scoring across different AFLP gels of the same primer combination, samples of two control genotypes were repeated every 4 lanes to standardize band scoring within and between autoradiographs of the same primer combination. Control genotypes used were NCo376 and R570. NCo376 was chosen as it occurs in the pedigree of many of the SASRI derived genotypes, and is expected to share many bands in common. R570 was chosen as the second control as several studies have been conducted with this genotype, and a reference map consisting of both RFLP and AFLP markers is available. (e.g. Grivet *et al.*, 1996, Hoarau *et al.*, 2001). Polymorphic fragments were scored as present (1) or absent (0) with a naming system that used the primer code from Table 3.4 to indicate the respective *Eco*R1 and *Mse*1 selective nucleotides, followed by a sequential number. For example, marker 1.2.A01 is the first polymorphism scored from primer combination AAC-CAC.

3.2.4. Population stratification.

Population admixture may cause false associations to be 'detected' due to allele frequency differences between sub-populations (Lander and Schork, 1994, Pritchard *et al.*, 2000.). In order test for population stratification, genetic distance between all genotypes was calculated using the Dice index,

$$Dij = \frac{b+c}{2a+b+c}$$

where a = the number of characters present in *i* and *j*;
b = the number of characters present in *i* and absent in *j*;
c = the number of characters absent in *i* and present in *j*.

The resulting matrix was used to build a neighbour-joining (N-J) tree using the DARwin 4.0 software package (Perrier *et al.*, 2003).

Potential mild population stratification suggested by the derived tree (see Results) was investigated further. The genotypes were divided into two groups based on the N-J tree, and marker frequency differences across 1053 loci between the two groups were compared using a chi-squared test (Prichard and Rosenberg, 1999). This was compared against the chi-square distribution, and also against a bootstrap random sampling of the population. Two groups of the same size as the populations suggested from the N-J tree (25 and 48 individuals respectively) were sampled with replacement from the dataset for 100 000 cycles, and the mean and standard error of the individual chi-squared values were calculated. This was then compared against the value obtained for the potentially sub-divided populations. Routines for the bootstrap analysis were programmed within the GAUSS™ 7.0 mathematical and statistical system language.

3.2.5. Marker identification and ideotype construction.

In order to identify markers associated with smut and eldana resistance ratings, Pearson's correlation coefficient was calculated for all markers with a frequency between 0.1 and 0.9, and for which there was less than five missing marker data-points. Smut rating and eldana rating were used as the dependent variables. This was done in Microsoft Excel. In addition, because phenotype for quantitative traits will be influenced by several to many loci, as well as by interactions between QTA's, stepwise multiple regressions was done to assemble sets of six markers ascribing the maximum amount of phenotypic variation in resistance score. With the multiple regression analysis, marker combinations were restricted to those having five or less missing values. In order to perform the large number of analyses required,

custom routines were written in GAUSS™ 7.0. This allowed flexibility in choosing which markers to add to the statistical models, as the correlation between individual markers with both phenotypic traits could be taken into account. The results of the multiple regression was used to derive marker ideotypes for groups of six markers associated with either resistance or susceptibility. The resistance ideotype will depict the presence of markers associated with resistance and the absence of markers associated with susceptibility – e.g. 111100 implies that the first four markers in the regression model are associated with resistance, and the last two associated with susceptibility. The ideotype for susceptibility will be the inverse, i.e. 000011, indicating the absence of the resistance markers and the presence of those conferring susceptibility. Predicted phenotypic score was calculated from the ideotypes using the partial regression coefficients from the multiple regression models.

3.2.6. Ideotype prediction for progeny of different possible cross combinations.

The goal of a gene/allele pyramiding strategy (GAP) is to produce progeny containing combinations of desirable alleles based on those present in the parent population. Data on the presence/absence of the markers involved in the ideotypes derived from section 3.2.5 above was extracted for each individual in the population. For 77 individuals, there are (77 x 76) ÷ 2 = 2926 possible bi-parental combinations, excluding selfs. The marker information for each individual was treated as a vector, and summed for all possible bi-parental combinations, giving a cross vector retaining information on marker dosage for the cross. For example, a potential cross between Genotype A (marker vector 100110) and Genotype B (marker vector 110010) will give a cross vector of 210120, indicating that both parents have the presence of Marker 1, one parent has Marker 2 present, both have Marker 3 absent, etc. The proportion of progeny having each marker will depend on the dosage – i.e. whether it is present in one or both parents, as well as the number of copies present within each parent. For ease of calculation, markers present in an individual were assumed to be single-copy – i.e. present on one chromosome only. In that case, a cross vector value of 1 implies that 50% of progeny will inherit the marker, and a cross vector value of 2 implies that 75% will inherit the marker (50% will have one copy, and 25% will have 2 copies). Each marker *present* (i.e. not absent) in the cross vector will segregate to give 2^x possible progeny vectors, where x is the number of markers present. For the example given of cross vector 210120, four markers are present and will segregate, giving 16 possible progeny marker vectors. The proportion of progeny expected for each vector class will depend on the cross vector value – i.e. present in one or both parents. This can be calculated from $x - 1$ sequential contingency tables, starting from a 2 x 2 table (Marker1 vs Marker2), followed by vectoring each preceding matrix of marker combinations and multiplying it by the next marker vector (4 x 2, M1M2 vs Marker3, then 8 x 2, M1M2M3 vs Marker4 etc). This was done for all

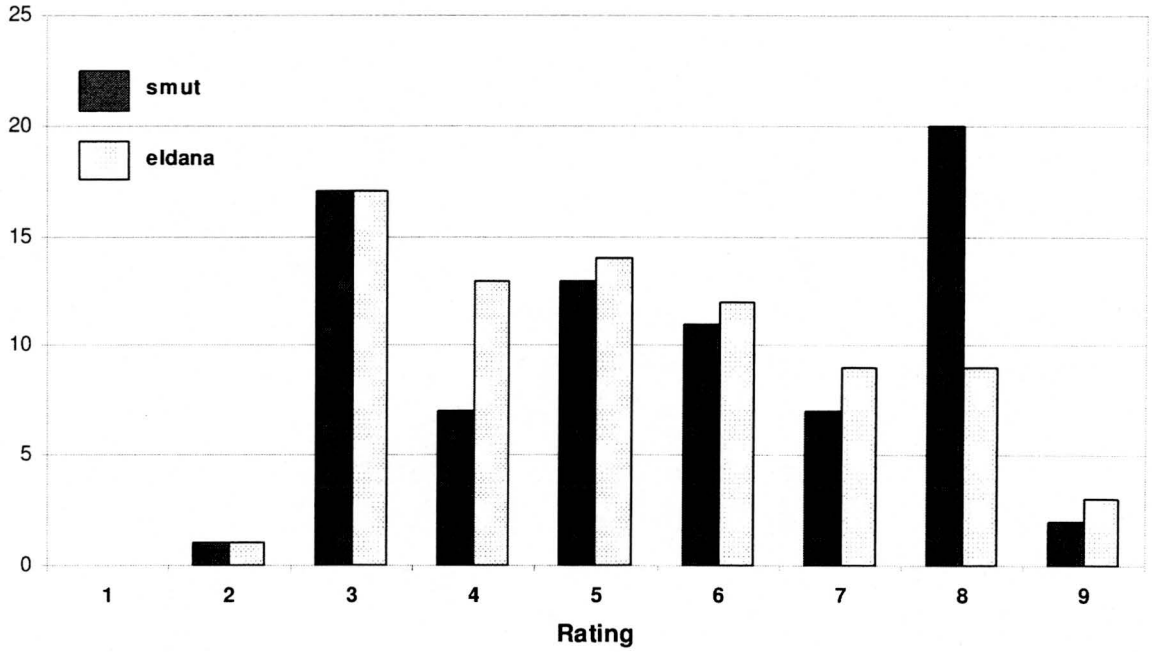
parent combinations of interest for the marker ideotypes derived from 3.2.5 above. For each possible progeny marker vector, predicted phenotypic rating was calculated from the partial regression coefficients. Potential crosses that result in pyramiding the desirable markers into progeny at a high frequency could then be identified, and targeted as priorities in breeding.

3.3. Results

3.3.1. Phenotypic data and marker-typing.

The distribution of phenotypic resistance ratings on a scale of 1 to 9 for both smut and eldana is shown in Figure 3.1. Correlating smut rating versus eldana rating produced a regression coefficient of r = -0.39 with a t-statistic of -3.37 (P = 0.0012, df = 76), indicating a moderate but highly significant negative association between smut and eldana resistance within this data set.

Figure 3.1. Distribution of phenotypic rating for smut and eldana across the population of 77 genotypes. 1 indicates highly resistant, and 9 highly susceptible.



Southern analysis using 45 RFLP probes and 2 restriction enzymes produced 275 polymorphic markers, with a frequency ranging from 0.08 to 0.94 and an average of 3.34 bands per probe/enzyme combination. Bands of different size produced by the same probe/enzyme combination can either be different alleles at the same locus, or represent genetic variation across duplicated loci. Of the 64 AFLP primer combinations used, 13

produced banding profiles that were either highly monomorphic or difficult to score. The remaining 51 score-able combinations yielded 1056 polymorphic markers with a frequency ranging from 0.04 to 0.96 and an average of 20.7 scored fragments per combination. The distribution of markers across frequency bin ranges for both RFLP and AFLP markers is shown in Table 3.5.

Table 3.5. Numbers of RFLP and AFLP markers at different frequencies within the population

Frequency	Number of markers	
bin	RFLP	AFLP
0 - 0.10	5	15
0.11 - 0.20	17	108
0.21 - 0.30	42	180
0.31 - 0.40	55	169
0.41 - 0.50	32	190
0.51 - 0.60	31	145
0.61 - 0.70	41	123
0.71 - 0.80	29	85
0.81 - 0.90	19	35
0.91 - 1	4	7
Total	275	1057

Data checking revealed that the genotype 82F2907 was almost identical in marker profile to the variety NCo376. Fingerprinting the germplasm collection from which the DNA stocks had been sampled using microsatellite markers confirmed that the accession 82F2907 in the collection was indeed NCo376, and was a labeling error in the field. This genotype was removed from further analyses. It did, however, provide an opportunity to estimate the unbiased rate of scoring errors. Of the 1056 AFLP fragments scored, 42 were inconsistent between NCo376 and the sample labeled as 82F2907, suggesting a scoring error rate of ~4%. Experience at CIRAD with AFLPs in sugarcane (Hoarau *et al.* 2001; Asnaghi *et al.* 2004; Raboin *et al.* 2006) suggests that genotyping errors, i.e. the frequency of bands that would be scored differently between two repetitions of the same genotype is in the range of one percent (Raboin, personal communication[1]). That estimate, however, is derived from studies of a selfed progeny population, where DNA fragments segregate in predictable frequencies, with a minimum frequency of 0.75 within the population. Examining the miss-scored markers in this population of parent genotypes showed that 71% (30/42) had a frequency lower than 0.4, with the remaining 29% (12/42) having a frequency between 0.4

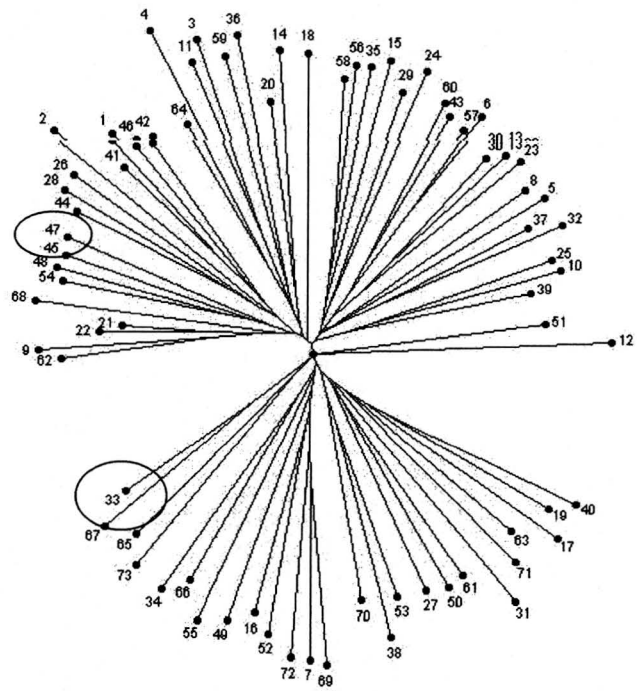
[1] Louis Marie Raboin, CIRAD. Currently rice breeder at CIRAD, Madagascar. Email: raboin@cirag.mg

and 0.5. As it can be difficult to align the position of low frequency markers across electrophoresis gels, it is not unexpected that these markers have higher scoring error rates. If only the markers with frequencies greater than 0.4 are considered, the scoring error rate drops to 1.1%, consistent with that observed by Raboin *et al.* (2006).

3.3.2. Population stratification.

The neighbor-joining tree representation of genetic diversity between individuals constructed from 1053 AFLP markers showed an even 'bicycle spoke' distribution (Figure 3.2), similar to that desirable in an ideal population for association mapping (Figure 2a from Yu and Buckler, 2006). A slight discontinuity in the pattern, however, similar to that depicted in Figure 2c from Yu and Buckler (2006), suggested that the population might be stratified into two sub-groups of 25 and 48 individuals respectively.

Figure 3.2. Neighbour-joining tree representing the genetic diversity at 1053 AFLP loci for the marker identification population of 77 genotypes. Diversity analysis and tree construction was done using DARwin 4.0 (Perrier *et al.,* 2003). Circled genotypes are the full-sibs NCo376 (33) and NCo339 (47).



In order to test this further, marker frequencies were calculated for all markers in the two potential groups, and subjected to chi-squared analysis to determine if marker frequencies were different in the two populations. For additional verification 100 000 bootstrap samples

of 25 and 48 individuals were taken from the dataset with replacement, chi-square analysis was done and the mean and standard error of the chi-square values calculated. The χ^2 value for marker frequency differences between the two sub-populations was 30.2, which is non-significant for the population size used. The mean and standard deviation for 100 000 bootstrap samples was 35.7 and 7.22 respectively. These results indicate that the population stratification suggested in Figure 3.2 is not significant, and that the dataset can be analyzed as individuals from a single population, with minimal risk of detecting false associations. Additional empirical evidence to support this is that individuals with known pedigree relationships such as full-sibs and parent-offspring, verified through molecular analyses occur in both sections of the potential sub-groups indicated in Figure 3.2. For example, the genotypes labeled as 33 and 47 and circled on the tree are varieties NCo376 and NCo339, which are full-sibs from the same cross.

3.3.3. Marker identification.

For convenience, a correlation coefficient threshold was set at |0.25|, equivalent to an R^2 value of 6.25%, which will be significant at P = 0.05 for the population size and range of marker frequencies present in the data set. A Bonferroni correction for significance thresholds for large numbers of independent variables (Bonferroni, 1936) was not applied, as it was of more interest to gain a general picture of the existence of potential markers throughout the genome, than in controlling the Type 1 error rate. The number of markers associated with either resistance or susceptibility for both AFLP and RFLP is shown in Table 3.6.

Table 3.6. Number of AFLP and RFLP markers associated with resistance or susceptibility to smut and eldana at r > |0.25|.

	Number of markers		Sub-	
	Resistant	**Susceptible**	**total**	**Total**
smut				
AFLP	39	13	52	
RFLP	7	5	12	64
eldana				
AFLP	55	41	96	
RFLP	13	6	19	115

Sixty four markers were associated with smut rating, and 115 with eldana rating. The individual markers with the largest effect ascribed 15.9% and 20.2% of the phenotypic variation in smut and eldana rating respectively. The strongest six markers for each trait are

given in Table 3.7. Due to the rating system used, a negative correlation implies association with resistance (i.e. low rating value), while a positive correlation implies association with susceptibility (i.e. high rating value). For smut, five of the strongest six markers were associated with resistance, and one with susceptibility. Two markers, *viz.* 7.8.B01 and 7.2.E02 showed some association with eldana phenotype as well. For eldana, four of the strongest six markers were associated with resistance, and two with susceptibility. One marker (6.6.A03) showed significant correlation with smut phenotype as well, and two other, *viz* 4.7.B04 and CelSynH1 showed some association with smut.

Table 3.7. Strongest six markers associated with smut and eldana respectively. Their correlation with the alternative trait is also given

	Correlation coefficient		
	smut	**eldana**	
smut			**R^2 - smut**
1.1.D02	-0.40	0.00	0.159
6.7.D01	-0.40	0.08	0.158
8.4.B03	-0.39	0.12	0.151
7.8.B01	-0.37	0.22	0.140
4.2.C03	-0.37	-0.01	0.140
7.2.E02	0.35	-0.21	0.122
eldana			**R^2 - eldana**
3.6.B02	0.02	-0.45	0.202
6.6.A03	-0.26	0.42	0.180
8.1.E03	0.14	-0.40	0.156
4.7.B04	-0.20	0.39	0.152
3.7.C01	0.06	-0.38	0.148
CelSynH1	0.19	-0.38	0.147

Five markers were significantly associated with both smut and eldana but all were of opposite signs, implying they are associated with QTLs involved in the negative phenotypic correlation observed between smut and eldana ratings. These are shown in Table 3.8. Four of these are anonymous AFLP markers, but one is derived from an RFLP probe with homology to the *Pto* gene encoding a serine/threonine kinase previously reported to be involved in resistance reaction in tomato to bacterial speck disease caused by *Pseudomonas syringa*e pv. *tomato* (Loh and Martin, 1995).

Table 3.8. Markers significantly with both smut and eldana. A positive correlation indicates susceptibility, and a negative sign, resistance

	Correlation coefficient	
Marker	**smut**	**eldana**
2.6.A05	0.275	-0.286
6.3.C05	0.256	-0.264
6.6.A03	-0.261	0.425
8.3.C09	-0.277	0.251
PtoKinH6	0.263	-0.355

In order to illustrate possible allelic variation at putative QTLs involved in resistance, results for nine markers scored from the same RFLP probe/enzyme combination are shown in Table 3.9. These markers were derived from an EST with homology to a peroxidase enzyme. A defense response known as the oxidative burst that involves the production of potentially toxic amounts of H_2O_2 and O_2^- has been associated with plant/pathogen interaction in several species (Legendre *et al.*, 1993) and has been linked with pathogen response elicited by the *Pto* gene in tomato (Chandra *et al.*, 1996).

Table 3.9. Correlation coefficients (r) and the size of the effect for putative allelic RFLP markers associated with smut and eldana, along with their frequencies within the population. The RFLP probe has homology to a peroxidase, and scored fragments were derived by digestion with either *Dra*1 (D) or *Hind*III (H). Significant associations are shown in bold.

	Correlation coefficient		Marker effect		Marker
Marker	**smut**	**eldana**	**smut**	**eldana**	**frequency**
PerOxD1	-0.20	**0.28**	-1.28	1.70	0.89
PerOxH1	0.03	**-0.27**	0.14	-1.16	0.24
PerOxH2	0.00	-0.05	0.02	-0.19	0.49
PerOxH3	-0.19	0.03	-1.26	0.18	0.10
PerOxH4	**0.28**	-0.19	1.11	-0.69	0.54
PerOxH5	-0.13	**0.25**	-0.51	0.93	0.44
PerOxH6	0.09	-0.21	0.61	-1.35	0.10

Peroxidases are one of the enzyme classes, along with catalases, superoxide dismutases and glutathione, which are thought to be involved in the oxidative burst (Legendre *et al.*, 1993). Three of the peroxidase derived markers are associated with reaction to eldana; one associated with resistance (PeroxH1), and two with susceptibility (PeroxD1 and PeroxH5). A third marker, PeroxH4 is significantly associated with susceptibility to smut. PeroxH3 and PeroxH6 have some association with resistance to smut and eldana at a probability level of $P = 0.11$ and $P = 0.07$ respectively. PeroxH2 has no association with either smut or eldana.

3.3.4. Multiple regression analysis for ideotype derivation.

Stepwise regression was done to identify sets of six markers ascribing the maximum amount of variation in each trait. Table 3.10a shows the results for smut, chosen by ignoring the associated effect of the marker on eldana. Table 3.10b shows an alternative set of six markers for smut, assembled by choosing the strongest markers associated with smut that were not also negatively correlated with eldana score.

Table 3.10. Stepwise regression for markers associated with smut. The corresponding effect of the same markers on eldana is also shown. The predicted phenotypic score for the resistant and susceptible ideotype is given for both traits. **3.10a** is the set chosen while ignoring the associated correlation with eldana. **3.10b** is the set chosen by excluding markers having a negative correlation with eldana. R and S refer to the resistant and susceptible ideotype respectively.

a.

	SMUT		ELDANA	
	Effect	**Prob.**	**Effect**	**Prob.**
Constant	5.56	<0.0001	5.24	<0.0001
7.8.B01	-1.47	<0.0001	0.84	0.065
8.6.C05	-1.62	<0.0001	1.03	0.022
4.6.D03	-2.28	<0.0001	1.53	0.003
3.8.D03	1.40	0.001	-0.75	0.128
8.5.C06	1.31	0.001	-1.19	0.007
7.2.D06	1.10	0.002	-0.40	0.329
R^2	0.553		0.252	
F-value	13.59	<0.0001	3.70	0.003
Ideotype				
111000 (R)	0.2		8.6	
000111 (S)	9.4		2.9	

b.

	SMUT		ELDANA	
	Effect	**Prob.**	**Effect**	**Prob.**
Constant	6.77	<0.0001	5.29	<0.0001
1.1.D02	-1.56	0.001	-0.02	0.972
4.7.E03	1.34	<0.0001	-0.14	0.783
FatRedH3	-1.00	0.007	-0.16	0.754
6.2.E03	2.45	0.001	0.07	0.943
8.2.C01	-1.37	0.004	0.04	0.945
4.2.C07	-1.09	0.003	0.16	0.751
R^2	0.549		0.005	
F-value	13.19	<0.0001	0.05	0.999
Ideotype				
101011 (R)	1.7		5.3	
010100 (S)	10.6		5.2	

The strongest set of six markers ascribed 55.3% of the phenotypic variation in smut rating. This was highly significant, with the F-value due to regression having a probability of P >0.0001. The first three markers in 3.10a have a negative effect in smut rating and are associated with resistance. The last three markers have a positive effect on smut rating, and are associated with susceptibility. The predicted ideotype for resistance – i.e. the presence of the resistance markers and the absence of the markers for susceptibility – gives a predicted rating of 0.2, while that for susceptibility give a prediction of 9.4. Due to the correlated effect of these markers with eldana phenotype, however, this set of markers was also significantly associated with susceptibility to eldana (P = 0.003), ascribing 25.2% of the variation in rating. The ideotype for smut resistance gave a predicted eldana rating of 8.6. Using this set of markers for breeding for smut resistance would therefore result in an undesirable correlated selection response, and cause an increase in eldana susceptibility.

Repeating the multiple regression and excluding markers showing a strong negative correlation between the two traits resulted in the marker set shown in Table 3.10b. This set of markers explains 54.9% of the phenotypic variation in smut rating, with an F-value and significance probability comparable to that shown in Table 3.10a. These markers, however, are not significantly correlated with eldana rating ($P = 0.999$), and do not result in an increase in eldana susceptibility, as seen by comparing the predicted rating for the ideotypes associated with smut resistance and susceptibility, viz. 5.3 versus 5.2.

Similar analyses for eldana are shown in Tables 3.11a and 3.11b. For this trait, selecting markers ignoring their effect on smut explains 63.4% of the variation in eldana score ($P < 0.0001$), and 20.5% of the variation in smut score ($P = 0.016$). As in the case with smut, selecting on the markers from Table 3.11a results in a predicted resistance to eldana, accompanied by a correlated predicted susceptibility to smut. When the correlation between the two traits is taken into account, a second set of markers still explains 61.5% of the variation in eldana score, without having a negative correlated effect on smut (Table 3.11b).

Table 3.11. Stepwise regression for markers associated with eldana. The corresponding effect of the same markers on smut is also shown. The predicted phenotypic score for the resistant and susceptible ideotype is given for both traits. **3.11a** is the set chosen while ignoring the associated correlation with smut. **3.11b** is the set chosen by excluding markers having a negative correlation with smut. R and S refer to the resistant and susceptible ideotype respectively.

a.

	ELDANA		SMUT	
	Effect	**Prob.**	**Effect**	**Prob.**
Constant	4.09	<0.0001	6.50	<0.0001
6.6.A03	1.96	<0.0001	-1.43	0.005
CelSynH1	-1.63	<0.0001	0.83	0.120
6.2.E05	-1.03	0.001	0.72	0.114
2.3.A02	1.69	<0.0001	-1.33	0.046
6.3.E01	-1.54	<0.0001	0.42	0.460
8.6.E01	1.19	<0.0001	-0.10	0.837
R^2	0.634		0.205	
F-value	19.08	<0.0001	2.84	0.016
Ideotype				
011010 (R)	-0.1		8.5	
100101 (S)	8.9		3.7	

b.

	ELDANA		SMUT	
	Effect	**Prob.**	**Effect**	**Prob.**
Constant	5.77	<0.0001	5.30	<0.0001
3.6.B02	-1.36	<0.0001	0.02	0.968
4.6.A01	1.40	<0.0001	-0.34	0.593
6.7.A03	1.66	<0.0001	-0.22	0.677
Pho301H2	-0.78	0.008	0.06	0.903
5.7.E04	-1.20	0.009	-0.19	0.813
3.8.D05	-0.78	0.011	0.71	0.180
R^2	0.615		0.036	
F-value	17.32	<0.0001	0.409	0.871
Ideotype				
100111 (R)	1.7		5.9	
011000 (S)	8.8		4.7	

3.3.5. *Selecting parent combinations expected to give desirable marker ideotypes in progeny.*

For sets of six markers scored as zero or one, $2^6 = 64$ different marker vectors are possible. From the markers and regression models in Tables 3.10b and 3.11b, predicted phenotype was calculated for all 64 possible marker vectors, and the five combinations with the lowest predicted score were selected. The dataset was then queried to extract those genotypes having the desirable ideotypes. The results for smut are shown in Table 3.12. Only one genotype, 78F0909, had the complete resistance ideotype of 101011. Five other genotypes had one of the partial resistance ideotypes shown in Table 3.12.

Table 3.12. Five resistance ideotypes with the lowest predicted smut rating, along with the genotypes within the population having one of these ideotypes.

Markers						Predicted
1.1.D02	**4.7.E03**	**FatRedH3**	**6.2.E03**	**8.2.C01**	**4.2.C07**	**rating**
1	0	1	0	1	1	1.7
1	0	0	0	1	1	2.7
1	0	1	0	1	0	2.8
1	1	1	0	1	1	3.1
1	0	1	0	0	1	3.1
						Genotype
1	0	1	0	1	1	78F0909
1	0	1	0	1	0	N22
1	0	1	0	1	0	N52/219
1	0	1	0	1	0	87L0329
1	0	1	0	0	1	CB38-22

The results for eldana are shown in Table 3.13. Again, only one genotype, CB38-22, had the complete resistance ideotype, with six others having partial resistance ideotypes. The fact that the resistance ideotypes for both traits are rare in individuals within the population illustrates the potential advantage of using a GAP strategy to pyramid the desirable markers in progeny through making specific bi-parental combinations predicted to result in the resistance ideotypes.

In order to identify the combinations resulting in the pyramiding of markers into the desired ideotypes, the marker profiles of all 2926 possible parent combinations between the 78 genotypes were compared. Cross combinations or cross marker vectors giving the complete resistance ideotype for either smut or eldana were extracted.

Table 3.13. Five resistance ideotypes with the lowest predicted eldana rating, along with the genotypes within the population having one of these ideotypes.

Markers						Predicted
3.6.B02	4.6.A01	6.7.A03	Pho301H2	5.7.E04	3.8.D05	rating
1	0	0	1	1	1	1.7
1	0	0	0	1	1	2.4
1	0	0	1	1	0	2.4
1	0	0	1	0	1	2.9
0	0	0	1	1	1	3.0
						Genotype
1	0	0	1	1	1	CB38-22
1	0	0	0	1	1	76M1566
1	0	0	1	0	1	N21
1	0	0	1	0	1	B42231
1	0	0	1	0	1	N20
1	0	0	1	0	1	80M1257
0	0	0	1	1	1	N33

For smut, 119 parent combinations resulted in crosses with the predicted resistance ideotype. Due to the fact that the marker combinations explain only 54.9% of the phenotypic variation in smut score, some of these combinations were between parents that had a smut rating of 5 or greater, indicating phenotypic susceptibility to smut. Filtering the list for combinations between parents that both had smut rating less than 5 gave a reduced list of 47 cross combinations between resistant parents resulting in crosses that would produce the resistant ideotype. These cross combinations involved 19 different parent genotypes. If phenotype alone was the only criterion for choosing parent combinations, 300 crosses are possible between parents both having smut ratings of 4 or less. The use of marker information has therefore allowed this large set to be narrowed down to those crosses which may be expected to have a greater chance of producing offspring with desirable resistance characteristics.

The list of 47 different cross combinations with their cross marker vectors is given in Table 3.14. The smut resistance ideotype from Table 3.10b used for this analysis has four resistance markers present. Each marker will segregate in the progeny, giving 16 different possible progeny marker vectors, or ideotype classes. Each ideotype class will have a different predicted smut rating. The 16 possible classes, with their predicted rating for smut are shown in Table 3.15.

Table 3.14. Parental cross combination resulting in the smut resistance ideotype. A value of 2 indicates that both parents have the marker present, while 1 indicates presence in one parent only. The average predicted rating, weighted across the respective expected ideotypes is given, as well as the percentage of progeny expected to be resistant, with a rating less than 3.5. M1 - M6 are the markers shown in Table 3.12.

Parent 1	Parent 2	Cross marker vector						Weighted rating	% progeny <3.5
		M1	M2	M3	M4	M5	M6		
78F0909	N22	2	0	2	0	2	1	3.3	63.3
78F0909	N52/219	2	0	2	0	2	1	3.3	63.3
78F0909	87L0329	2	0	2	0	2	1	3.3	63.3
N23	78F0909	1	0	2	0	2	1	3.3	51.6
CB38-22	N22	2	0	2	0	1	1	3.6	51.6
CB38-22	N52/219	2	0	2	0	1	1	3.6	51.6
CB38-22	87L0329	2	0	2	0	1	1	3.6	51.6
78F0909	CB38-22	2	0	2	0	1	2	2.9	49.2
78F0909	82F0675	2	0	1	0	1	2	3.3	42.2
78F0909	CB40/35	1	0	2	0	1	2	3.4	42.2
78F0909	85F1628	1	0	2	0	1	2	3.4	42.2
78F0909	N24	1	0	2	0	1	2	3.4	42.2
78F0909	84F2753	1	0	2	0	1	2	3.4	42.2
78F0909	N30	1	0	2	0	1	2	3.4	42.2
78F0909	N26	1	0	2	0	1	2	3.4	42.2
N22	82F0675	2	0	1	0	1	1	3.9	40.6
82F0675	N52/219	2	0	1	0	1	1	3.9	40.6
82F0675	87L0329	2	0	1	0	1	1	3.9	40.6
N23	CB38-22	1	0	2	0	1	1	4.0	40.6
78F0909	85L1056	1	0	2	0	1	1	4.0	40.6
78F0909	NCo339	1	0	2	0	1	1	4.0	40.6
78F0909	82F2907	1	0	2	0	1	1	4.0	40.6
78F0909	80L0432	1	0	2	0	1	1	4.0	40.6
78F0909	N11	1	0	2	0	1	1	4.0	40.6
CB40/35	N22	1	0	2	0	1	1	4.0	40.6
CB40/35	N52/219	1	0	2	0	1	1	4.0	40.6
CB40/35	87L0329	1	0	2	0	1	1	4.0	40.6
85F1628	N22	1	0	2	0	1	1	4.0	40.6
85F1628	N52/219	1	0	2	0	1	1	4.0	40.6
85F1628	87L0329	1	0	2	0	1	1	4.0	40.6
N24	N22	1	0	2	0	1	1	4.0	40.6
N24	N52/219	1	0	2	0	1	1	4.0	40.6
N24	87L0329	1	0	2	0	1	1	4.0	40.6
N22	84F2753	1	0	2	0	1	1	4.0	40.6
N22	N30	1	0	2	0	1	1	4.0	40.6
N22	N26	1	0	2	0	1	1	4.0	40.6
84F2753	N52/219	1	0	2	0	1	1	4.0	40.6
84F2753	87L0329	1	0	2	0	1	1	4.0	40.6
N52/219	N30	1	0	2	0	1	1	4.0	40.6
N52/219	N26	1	0	2	0	1	1	4.0	40.6
N30	87L0329	1	0	2	0	1	1	4.0	40.6
87L0329	N26	1	0	2	0	1	1	4.0	40.6
78F0909	J59/3	1	0	1	0	1	2	3.8	34.4
N23	82F0675	1	0	1	0	1	1	4.3	31.3
J59/3	N22	1	0	1	0	1	1	4.3	31.3
J59/3	N52/219	1	0	1	0	1	1	4.3	31.3
J59/3	87L0329	1	0	1	0	1	1	4.3	31.3

Table 3.15. Progeny ideotypes possible from the parent combinations shown in Table 3.14, along with their predicted smut rating.

Possible progeny ideotypes						Predicted
M1	M2	M3	M4	M5	M6	rating
1	0	1	0	1	1	1.75
1	0	0	0	1	1	2.75
1	0	1	0	1	0	2.84
1	0	1	0	0	1	3.11
0	0	1	0	1	1	3.30
1	0	0	0	1	0	3.84
1	0	0	0	0	1	4.12
1	0	1	0	0	0	4.21
0	0	0	0	1	1	4.31
0	0	1	0	1	0	4.40
0	0	1	0	0	1	4.67
1	0	0	0	0	0	5.21
0	0	0	0	1	0	5.40
0	0	0	0	0	1	5.67
0	0	1	0	0	0	5.76
0	0	0	0	0	0	6.77

The 47 cross combinations shown in Table 3.14 are represented by ten different cross marker vectors. For each potential cross, the proportion of progeny expected in each of the 16 possible ideotype combinations was calculated from the parent combination marker value (i.e. 1 or 2, depending on whether one or both parents had the marker present). The predicted smut rating was calculated for each ideotype, and the average rating, weighted over the proportion of progeny expected in each class was calculated. This is also shown in Table 3.14. In addition, the proportion of progeny in expected resistance ideotype classes with a predicted smut rating less than 3.5 was calculated, and shown in Table 3.14. The crosses are shown ranked by percent progeny with rating less than 4.5, followed by weighted average cross rating.

The first three cross combinations in Table 3.14 involving the genotypes 78F909, N22, N52/219 and 87L0329 all share the same cross marker vector of 202021, indicating that the first three resistance markers, are present in both parents, while the fourth resistance marker is present in one parent only. The expected segregation to the 16 possible progeny ideotypes, combined with the predicted rating for each ideotype (from Table 3.10b), results in a prediction that 63% of the progeny derived from such a cross would have a marker ideotype with a resistant rating prediction less than 3.5, and a weighted average smut rating across all possible ideotypes of 3.3. The four crosses at the bottom of Table 3.14 share the

cross marker vector of 101011, indicating that these are complementary combinations, with each resistance marker being present in only one of the parents. The expected segregation of this cross ideotype to the 16 possible progeny ideotypes predicts that only 31% of the progeny would fall in ideotype classes with a predicted rating of less than 3.5, with a weighted average across all ideotypes of 4.3. In terms of breeding for smut resistance, priority would obviously be given to those cross combinations resulting in a high predicted number of progeny with a smut rating of less than 3.5, and a low weighted cross mean rating.

The same exercise was repeated for the eldana resistance ideotype. Cross combinations resulting in progeny expected to have the resistance ideotype are shown in Table 3.16. Half of parent combinations are comprised of four different cross marker vector classes, *viz.* 200212, 200112, 100212 and 200111, which are predicted to have more than 50% of progeny with ideotypes resulting in a predicted eldana rating of less than 3.5. These would be desirable combinations to make in breeding for eldana resistance, as markers associated with resistance have been pyramided within the progeny at a relatively high frequency. If it is assumed that yield traits and sucrose content is un-correlated with eldana resistance, then selecting for high yield and high sucrose content within these crosses should result in some of the individuals selected also being resistant to eldana. In other words, selecting on yield traits that are easy to measure in the early stages of the selection programme should not result in the chance loss of eldana resistance, which should occur at high frequency within these families.

Four parent combinations resulted in the cross ideotype for both smut and eldana, and are shown in Table 3.17. These combinations afford the opportunity to use classical marker assisted selection to identify those individual progeny having the complete resistance ideotype for both traits. If such recombinant genotypes could be identified, they would be a valuable resource in resistance breeding for both traits. In most cases, however, the resistance markers are present in one parent only. Due to segregation, only a small percentage of progeny are expected to have the complete resistance ideotype for either smut or eldana. Multiplying the expected frequency for the individual traits gives the percentage of progeny expected to have the complete resistance ideotypes for both smut and eldana. In three of the combinations, less than 1% of progeny are expected to have the complete resistance ideotypes for both traits. Because of this low frequency, it may not be effective to try and identify those individuals with the full ideotype for both traits in these crosses. For the combination involving CB38-22 and 76M1566, almost 2% of the progeny are predicted to contain both ideotypes. This cross could potentially be used in classic MAS to identify the rare individuals having both resistance ideotypes.

Table 3.16. Parental cross combination resulting in the eldana resistance ideotype. A value of 2 indicates that both parents have the marker present, while 1 indicates presence in one parent only. The average predicted rating, weighted across the respective expected ideotypes is given, as well as the percentage of progeny expected to be resistant, with a rating less than 3.5. M1 - M6 are the markers shown in Table 3.13.

Parent 1	Parent 2	Cross marker vector						Weighted rating	% progeny <3.5
		M1	M2	M3	M4	M5	M6		
N21	CB38-22	2	0	0	2	1	2	2.6	66.0
B42231	CB38-22	2	0	0	2	1	2	2.6	66.0
CB38-22	N20	2	0	0	2	1	2	2.6	66.0
CB38-22	80M1257	2	0	0	2	1	2	2.6	66.0
N21	76M1566	2	0	0	1	1	2	2.9	56.0
B42231	76M1566	2	0	0	1	1	2	2.9	56.0
CB38-22	79F1855	2	0	0	1	1	2	2.9	56.0
N20	76M1566	2	0	0	1	1	2	2.9	56.0
76M1566	80M1257	2	0	0	1	1	2	2.9	56.0
N21	N33	1	0	0	2	1	2	3.1	53.1
B42231	N33	1	0	0	2	1	2	3.1	53.1
75L1463	CB38-22	1	0	0	2	1	2	3.1	53.1
81W447	CB38-22	1	0	0	2	1	2	3.1	53.1
N20	N33	1	0	0	2	1	2	3.1	53.1
80M1257	N33	1	0	0	2	1	2	3.1	53.1
85L1769	CB38-22	2	0	0	1	1	1	3.4	50.0
75L1463	76M1566	1	0	0	1	1	2	3.3	43.8
77F0637	CB38-22	1	0	0	1	1	2	3.3	43.8
79L1294	CB38-22	1	0	0	1	1	2	3.3	43.8
81W0447	76M1566	1	0	0	1	1	2	3.3	43.8
CB38-22	N19	1	0	0	1	1	2	3.3	43.8
CB38-22	87L1484	1	0	0	1	1	2	3.3	43.8
N33	79F1855	1	0	0	1	1	2	3.3	43.8
N17	CB38-22	1	0	0	2	1	1	3.5	43.8
CB38-22	80W1459	1	0	0	2	1	1	3.5	43.8
N17	76M1566	1	0	0	1	1	1	3.7	37.5
85L1769	N33	1	0	0	1	1	1	3.7	37.5
CB38-22	N22	1	0	0	1	1	1	3.7	37.5
80W1459	76M1566	1	0	0	1	1	1	3.7	37.5

Table 3.17. Parent combinations resulting in the predicted cross ideotype for both smut and eldana. The percentage of progeny with the complete resistance ideotype for smut and eldana, and the percentage of progeny falling into ideotype classes with predicted resistance for both traits is shown.

Parent 1	Parent 2	Smut	Eldana	Smut	Eldana	Both	Smut	Eldana	Both
		Cross vector		% with complete resistance ideotype (101011:10011)			% with predicted rating less than 3.5		
N23	CB38-22	102011	100111	9.38	6.25	0.59	40.6	37.5	15.2
CB38-22	76M1566	102011	200122	9.38	21.09	1.98	40.6	70.3	28.6
CB38-22	N22	202011	100111	14.06	6.25	0.88	51.6	37.5	19.3
76M1566	82F0675	101011	100111	6.25	6.25	0.39	31.3	37.5	11.7

If all possible ideotypes are considered, five different smut marker vectors result in a resistant prediction, with smut rating of less than 3.5. The proportion of resistant progeny expected from each cross ranges from 31% to 51%. For eldana, six different marker vectors result in resistance ideotypes, with predicted rating of less than 3.5. The proportion of eldana resistant progeny expected from each cross ranges from 37% to 70%. The individual ideotypes for both traits are shown in Table 3.18. If the two traits are combined, the cross between CB38-22 and 76M1566 is expected to have 28% of progeny resistant to both traits, while the combination between 76M1566 and 82F0675 is expected to give 11% of progeny resistant to both smut and eldana. The combined resistance is derived from all possible combinations of the five smut and six eldana resistance ideotypes, giving a total of 30 different marker vector combinations. Within-family selection for combined resistance would obviously be more efficient in the crosses with a higher predicted number of progeny with the desirable marker ideotypes.

Table 3.18. Resistance ideotypes with predicted rating of < 3.5 for either smut or eldana. Any of the 30 combinations possible should give combined resistance to both traits.

Smut resistance ideotypes						Predicted rating	Eldana resistance ideotypes						Predicted rating
1	0	1	0	1	1	1.7	1	0	0	1	1	1	1.7
1	0	0	0	1	1	2.7	1	0	0	0	1	1	2.4
1	0	1	0	1	0	2.8	1	0	0	1	1	0	2.4
1	0	1	0	0	1	3.1	1	0	0	1	0	1	2.9
0	0	1	0	1	1	3.3	0	0	0	1	1	1	3.0
							1	0	0	0	1	0	3.2

Selecting for yield and sucrose alone at early selection stages would likely result in the chance loss of the individuals containing the resistance ideotypes for both traits, as they are at low frequency within the progeny population. In order to prevent the loss of potentially valuable germplasm, progeny of these crosses could be screened with the 12 markers to identify individuals with one of the 30 different marker vector combinations possible. This will be more effective than trying to recover the very rare genotypes with the complete resistance ideotype. Genotypes identified with the desirable marker vectors will be a valuable resource in GAP breeding for combined resistance to smut and eldana.

3.4. Discussion.

Using an association or linkage disequilibrium approach, molecular markers associated with putative QTAs involved in the reaction of sugarcane to attack by eldana and smut have been identified. The phenotypic data on resistance score for both traits was derived from mining data from the SASRI database accumulated from trials assessed over different sites and seasons. One advantage of this approach is that QTAs identified are relatively free of site or season interaction effects, without incurring the considerable expense of planting specific trials over locations and years.

Individual markers identified through association analysis explained up to 20% of the variation in eldana resistance, and 16% of the variation in smut resistance (Table 3.7). In previous marker identification studies in sugarcane, Ming *et al.* (2002) reported on markers explaining 3.8% to 16.2% of the phenotypic variance in sugar yield, sucrose content, stalk weight, stalk number, fibre content and stalk ash content in two bi-parental crosses. Hoarau *et al*, (2002) identified markers describing 3% to 6% of the variation in brix, stalk length, stalk diameter and stalk number in a selfed population of the commercial variety R570. Aitken *et al.* (2006) found markers explaining 3% to 12% of the variation in sucrose content in a cross between a commercial cultivar Q165 and a *Saccharum officinarum* clone, IJ76-514.

In terms of markers for disease resistance, McIntyre *et al.*, (2005) identified seven markers explaining 4% to 16% of the variation in resistance to Pachymetra root rot, and four markers explaining 7% to 18% of the variation in brown rust in a bi-parental cross between two elite cultivars. When these markers were tested in a collection of 154 elite genotypes in an association approach, three of the Pachymetra resistance markers and one of the rust resistance markers remained significantly associated. The strength of marker-trait association reported from sugarcane is in a similar range found for disease traits in other polyploid crops. In cassava, Jorge *et al.*, (2000) identified markers explaining from 9% to 20% of the variation in resistance to bacterial blight disease in a segregating F1 population. In tetraploid potato, Simko *et al.*, (2004) identified a single marker that explained from 8% to 25% of the variation in resistance to Verticillium wilt disease in three different populations.

In terms of damage caused by insects, Groh *et al.* (1998) studied leaf-feeding damage of the southwestern corn borer (*Diatrea grandiosella*) and the sugarcane borer (*D. saccharalis*) in two maize populations. Individual markers explained between 2.1% and 25.8% of the variation in leaf damage. A set of nine markers explained 52.4% of the variation in damage due to the corn borer, and a set of 8 markers explained 52.8% of the variation in damage

caused by the sugarcane borer. The strength of the association between individual markers and resistance to smut and eldana reported in this chapter are thus similar to those found in other crops.

The number of markers explaining more than 6.25% of the variation in phenotype for the two traits was fairly high; 115 for eldana, and 64 for smut (Table 3.6). There may be several reasons for this. Firstly, due to the fact the marker identification was comprised of genotypes within the breeding programme, genetic variation for the traits of interest will be greater than that within the progeny of a single cross and it will possible to capture a wider diversity of alleles than is possible from a bi-parental marker discovery population. In the study of McIntyre *et al.* (2005) mentioned above, only seven and four significant markers were identified from a bi-parental cross for the two traits of interest, despite the fact that ~1000 markers were used. Table 3.5 shows that a large proportion of scored polymorphisms were at relatively low frequency. Of all markers, 28% had a frequency lower than 0.30 in the population; of the markers associated significantly with eldana and smut, 34% and 37% respectively were at a frequency of lower than 0.30. Many of these would not have been present in a bi-parental or selfed marker discovery population, and would have remained un-identified. The large numbers of markers identified for smut and eldana reported in this chapter illustrate the benefit of using an association approach within a breeding population for marker discovery. Although population stratification can lead to a high incidence of type 1 errors in detecting marker-trait associations within germplasm not derived from a bi-parental cross (Pritchard and Rosenberg, 1999), no evidence of sub-structure was seen in the population used in this study.

Some of these markers may represent allelic diversity at the same locus. Using RFLP markers, it was possible to identify markers that may represent different alleles (QTAs) at the same locus (QTL). Without a genetic map that can assign the markers to chromosomal locations, however, it is impossible to know whether the RFLP variants represent different alleles, or duplicated regions on non-homologous chromosomes. As it is known that many plant species including the grasses and cereals have undergone ancient genome duplication and rearrangement (e.g. Paterson *et al.*, 2004, Paterson *et al.*, 2005) it is likely that some loci with similar sequence are non-homologous. Dominantly scored AFLP fragments are anonymous in nature, so it is not possible to postulate any potential allelic relationships between them without a genetic map where homology relationships between linkage groups can be identified.

An additional explanation for the larger numbers of significant markers identified is that it is likely that some markers are linked on the same haplotype. Some QTAs may have several markers linked to them, so the number of functional alleles will be less than the number of marker-trait associations. Again, without a map showing linkage arrangements between markers, these relationships will remain unknown.

The number of genotypes in the marker discovery population will affect the statistical power of detecting significant associations. Beavis (1994) showed that in a simulated F_2 population, increasing the population size from 100 to 500 progeny increased the power of QTL detection from 33% to 86% for loci ascribing 6.3% of the phenotypic variation. It is likely that if the population size of this study had been increased from the 77 genotypes used, that some additional significant markers associations would have been detected. It is also likely, however, that new associations detected through increasing the population size would have smaller predicted effects on phenotype, and would not be used in practice. In addition, as the population size increases, the likelihood of population stratification existing also increases, which may result in an increase in the detection of false associations. For this study, the population used reflects a compromise between: (1) being small enough for manual RFLP and AFLP analysis, (2) large enough to detect markers with relatively large effects on phenotype, and (3) small enough to minimize the possibility of population stratification.

The identification of markers associated with the negative phenotypic correlation between smut and eldana (Table 3.8), and markers resulting in a negative correlated selection response (Tables 3.10a and 3.11a), is of significance in terms of the application of markers in breeding. Using markers associated with eldana resistance for selection without taking their association with smut into account would most likely result in an increase in smut susceptibility, and *visa versa*. It is possible, however, to choose sets of markers for resistance to one trait that are not correlated with the second trait (Tables 3.10b and 3.11b). For both smut and eldana, groups of six markers explained more than 50% of the phenotypic variation in resistance rating, with the individual effect of each marker being statistically significant. This is similar to the results reported by Groh *et al.* (1998) for borer damage in maize, where sets of 9 and 8 markers explained 52% of the variation in leaf damage due to corn borer and sugarcane borer. The magnitude of the effect found for smut and eldana implies that these markers can be used effectively in breeding to enrich the population for quantitative trait alleles involved in resistance, and to reduce the frequency of alleles associated with susceptibility using a GAP breeding strategy.

Using the resistance ideotypes derived from multiple regression, hypothetical bi-parental combinations can be made by summing the parent ideotypes across marker loci, in order to pyramid desirable marker sets in progeny populations. Progeny are expected to segregate at each marker locus, giving rise to a range of expected progeny ideotype classes. The proportion of progeny in each class is dependent on whether the marker is present in one or both parents, and this can be calculated. Cross combinations resulting in high predicted numbers of progeny in desirable ideotype classes can be identified, as shown in Tables 3.14 and 3.16. This appears to be a novel form of analysis and application of marker data, as equivalent reports have not been found in the literature to date for other crops. Identification of desirable combinations through cross vector analysis allows crossing efforts to be focused on those combinations that have a greater potential to give rise to resistant progeny. This is an important consideration in sugarcane breeding at SASRI, where the number of parent genotypes that can be used in crossing, as well as the number of cross combinations made, is limited by the capacity of the photoperiod and glasshouse facilities. The benefit of a GAP marker utilization strategy in this context is that by identifying cross combinations that give a high proportion of progeny expected to have resistant marker ideotypes, even random selection within these families will recover individuals with desirable marker ideotypes. In other words, it is not necessary to screen progeny populations using MAS, as the frequency of desirable marker combinations is high. Classical MAS may be useful in some cases, however, when desirable marker combinations are rare in the population and are considered important enough to justify the cost of screening many progeny, as illustrated in Table 3.17.

The fact that sugarcane is a polyploid complicates the application of markers in breeding. As multiple alleles can be present for the same QTL, knowing the simple presence or absence of an individual QTA may be misleading. For example, the phenotypic effect of the presence of one resistance allele may be nullified by the presence of multiple copies of alleles for susceptibility. If the allelic relationship between markers is known these effects can be taken into account, potentially increasing the power of using markers for breeding. In order to determine allelic markers at QTLs of interest, a genetic map is required. As conventional mapping populations based on segregating progeny from bi-parental or selfed crosses identify only a fraction of alleles present in the breeding population, an alternative strategy is required. A potential solution is to extend the association genetic approach used for marker identification as described in this chapter, in order to map entire populations using linkage disequilibrium methods. Linkage disequilibrium mapping has been pioneered within the context of human genetics (e.g. Reich *et al.*, 2001), is beginning to be applied in the context of plant genetics (e.g. Kraakman *et al.*, 2004, 2006, Breseghello and Sorrels, 2005) but has not yet been attempted in sugarcane. The remaining chapters of this thesis will focus on

describing the development of methodology for linkage disequilibrium mapping in sugarcane (Chapter 4), and the application of the methodology to the data set described here (Chapter 5). Any improvement resulting from a map-based approach to molecular breeding will then be used to refine the GAP strategy for breeding sugarcane for resistance to smut and eldana.

3.5. References.

Aitken, KS, Jackson, PA and McIntyre, CL. 2006. QTL identified for sugar related traits in a sugarcane (*Saccharum* spp.) cultivar x *S. officinarum* population. Theoretical and Applied Genetics 112: 1306-1317.

Asnaghi C., D. Roques, S. Ruffel, C. Kaye, J.Y. Hoarau, Telismart, H, Girard, JC, Raboin, LM, Risterucci, AM, Grivet, L and D'Hont, A. 2004. Targeted mapping of a sugarcane rust resistance gene (*Bru1*) using bulked segregant analysis and AFLP markers. Theoretical and Applied Genetics 108:759-764.

Beavis, WD. 1994. The power and deceit of QTL experiments: Lessons from comparative QTL studies. p. 250-266. In Wilkinson, D. (ed.). Proceedings of the 49th Annual Corn & Sorghum Research Conference, Chicago, IL. American Seed Trade Association, Washington, DC.

Beer, C, Siripoonwiwat, W, O'Donoughue, LS, Souza, E, Matthews, D, and M.E. Sorrells, ME. 1997. Associations between molecular markers and quantitative traits in an oat germplasm pool: Can we infer linkages? Journal of Agricultural Genomics. Published with permission from CAB International. Full text available from http://www.cabi-publishing.org/jag/papers97/paper197 /indexp197.html

Bonferroni, CE. 1936. Teoria statistica delle classi e calcolo delle probabilità. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze. 8: 3-62.

Breseghello, F and Sorrels, ME. 2005. Association mapping of kernel size and milling quality in wheat. Genetics 172: 1165-1177.

Butterfield, MK and Thomas, DW. 1996. Sucrose, yield and disease resistance characteristics of sugarcane varieties under test in the SASEX selection programme. Proceedings of the South African Sugar Technologists Association 70: 103-105.

Butterfield, MK, D'Hont, AD and Berding, N. 2001. The sugarcane genome: A synthesis of current understanding and lessons for breeding and biotechnology. Proceedings of the South African Sugar Technologists Association 75: 1-5.

Butterfield, MK, Rutherford, RS, Carson, DL and Huckett, BI. 2004. Application of gene discovery to varietal improvement in sugarcane. South African Journal of Botany 70: 167-172.

Chandra, S, Martin, GB and Low, PS. 1996. The *Pto* kinase mediates a signaling pathway leading to the oxidative burst in tomato. Proceedings of the National Academy of Sciences USA 93: 13393-13397.

Dellaporta, SL, Woods, J and Hicks, JB. 1983. A plant DNA minipreparation: Version II. Plant Molecular Biology Reporter 1: 19-21.

GAUSS Mathematical and Statistical System version 5. Aptech Systems Inc., Maple Vally, WA, 1984-2003.

Grivet, L, D'Hont, A, Roques, , Feldmann, P., Lanaud, C and Glaszmann, JC. 1996. RFLP mapping in cultivated sugarcane (*Saccharum* spp)–: Genome organization in a highly polyploid and aneuploid interspecific hybrid. Genetics 142: 987-1000.

Groh, S, Gonzalez-de-leon, D, Khairallah, MM, Jiang, C, Bergvinson, D, Bohn, M, Hoisington, DA and Melchinger, AE. 1998. QTL mapping in tropical maize III. Genomic regions for resistance to *Diatrea* spp. and associated traits in two RIL populations. Crop Science 38:1062-1072.

Heinze, BS, Thokoane, LN, Williams, NJ, Barnes, JM and Rutherford, RS. 2001. The smut-sugarcane interaction as a model system for the integration of marker discovery and gene isolation. Proceedings of the South African Sugar Technologists Association 75:88-93.

Hoarau, JY, Offmann, B, D'Hont, A, Risterucci, AM, Roques, D, Glaszmann, JC and Grivet, L. 2001. Genetic dissection of a modern cultivar (*Saccharum* spp.). I. Genome mapping with AFLP. Theoretical and Applied Genetics 103: 84-97.

Hoarau, JY, Grivet, L, Offmann, B, Raboin, LM, Diorflar, JP, Payet, J, Hellmann, M, D'Hont, A and Glaszmann, JC. 2002. Genetic dissection of a modern sugarcane cultivar (*Saccharum* spp.) II. Detection of QTLs for yield components. Theoretical and Applied Genetics 105: 1027-1037.

Hutchinson, PB. 1968. A note on disease resistance ratings for sugarcane varieties. Proceedings of the International Society of Sugar Cane Technologists 13: 1087-1089.

Igartua, E, Casas, AM, Ciudad, F, Montoya, JL and Romagosa, I. 1999. RFLP markers associated with major genes controlling heading date evaluated in a barley germ plasm pool. Heredity 83: 551-559.

Jorge, V, Fregene, MA, Duque, MC, Bonierbale, MW, Tohme, J and Verdier, V. 2000. Genetic mapping of resistance to bacterial blight disease in cassava (*Manihot esculenta* Crantz). Theoretical and Applied Genetics 101: 865-872.

Keeping, MG and Rutherford, RS. 2004. Resistance mechanisms of South African sugarcane to the stalk borer *Eldana saccharina* (Ledipodtera: Pyralidae). A review. Proceedings of the South African Sugar Technologists Association 78: 307-312.

Kraakman, ATW, Niks, RE, Van den Berg, PMMM, Stam, P and Eeuwijk, FA. 2004. Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars. Genetics 168: 435-446.

Kraakman, ATW, Martinez, F, Mussiraliev, B, van Eeuwijk, FA and Niks, RE. 2006. Linkage disequilibrium mapping of morphological, resistance and other agronomically relevant traits in modern spring barley cultivars. Molecular Breeding 17: 41-58.

Lander, ES and Schork, NJ. 1994. Genetic dissection of complex traits. Science 265: 2037-2048.

Legendre, L, Rueter, S, Heinstein, PF and Low, PS. 1993. Characterization of the oligogalacturonide-induced oxidative burst in cultured soybean (*Glycine max)* cells. Plant Physiology 102: 233-240.

Loh, YT and Martin, GB. 1995. The disease-resistance gene *Pto* and the fenthion-sensitivity gene *Fen* encode closely related functional protein kinases. Proceeding of the National Academy of Sciences USA 92: 4181-4184.

McIntyre, CL, Whan, VA, Croft, B, Magarey, R and Smith, GR. 2005. Identification and validation of molecular markers associated with Pachymetra root rot and brown rust in sugarcane using map and association based approaches. Molecular Breeding 16: 151-161.

Ming, R, Wang, YW, Draye, X, Moore, PH, Irvine, JE and Paterson, AH. 2002. Molecular dissection of complex traits in autopolyploids: mapping QTLs affecting sugar yield and related traits in sugarcane. Theoretical and Applied Genetics 105: 332-345.

Moore, PH and Nuss, KJ. 1987. Flowering and flower synchronization. Pp 102-127. In: Heinz, DJ (ed.) Sugarcane improvement through breeding. Elsevier, Amsterdam.

Paterson, AH, Bowers, JE and Chapman, BA. 2004. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. Proceeding of the National Academy of Sciences USA 101: 9903-9908.

Paterson, AH, Bowers, JE, Van de Peer, Y and Vandepoele, K. 2005. Ancient duplication of cereal genomes. New Phytologist 165: 658-661.

Perrier, X, Flori, A and Bonnet, F. 2003. Data analysis methods. In: Hamon, P, Seguin, M, Perrier, X and Glaszmann, JC. (eds). Genetic diversity of cultivated tropical plants. Enfield Science Publishers, Montpellier. pp 43-76.

Pritchard, JK and Rosenberg, NA. 1999. Use of unlinked genetic markers to detect population stratification in association studies. American Journal of Human Genetics 65: 220-228.

Pritchard, JK, Stephens, M, Rosenberg, NA and Donnelly, P. 2000. Association mapping in structured populations. American Journal of Human Genetics 67: 170-181.

Raboin, LM, Oliveira, KM, Lecunff, L, Telismart, H, Roques, D, Butterfield, M, Hoarau, JY and D'Hont, A. 2006. Genetic mapping in sugarcane, a high polyploidy, using bi-parental progeny: identification of a gene controlling stalk colour and a new rust resistance gene. Theoretical and Applied Genetics 112: 1382-1391.

Reich, D., Cargill, M, Bolk, S, Ireland, J, Sabeti, PC, Richter, DJ, Lavery, T, Kouyoumjian, R, Farhadian, SF, Ward, R and Lander, ES. 2001. Linkage disequilibrium in the human genome. Nature 411: 199-204

Simko, I, Costanzo, S, Haynes, K, Christ, BJ and Jones, RW. 2004. Linkage disequilibrium mapping of a *Verticillium dahliae* resistance quantitative trait locus in tetraploid potato

(*Solanum tuberosum*) through a candidate gene approach. Theoretical and Applied Genetics 108: 217-224.

Spoel, SH, Koornneef, A, Claessens, SMC, Korzelius, JP, Van Pelt, JA, Mueller, MJ, Buchala, AJ, Metraux, JP, Brown, R, Kazan, K, Van Loon, LC, Dong, X and Pieterse, CMJ. 2003. NPR1 modulates cross-talk between salicylate and jasmonate dependent defense pathways through a novel function in the cytosol. The Plant Cell 15: 760 – 770.

Thaler, JS, Karban, R, Ullman, DE, Boege, K and Bostock, RM. 2002. Cross-talk between jasmonate and salicylate plant defense pathways: effects on several plant parasites. Oecologia 131: 227-235.

Thokoane, LN and Rutherford, RS. 2001. cDNA-AFLP differential display of sugarcane (*Saccharum* spp. hybrids) genes induced by challenge with the fungal pathogen *Ustilago Scitaminea* (sugarcane smut). Proceedings of the South African Sugar Technologists Association 75:104-107.

Thornsberry, JM, Goodman, MM, Doebley, J, Kresovich, S and Nieksen, D. 2001. Dwarf8 polymorphisms associate with variation in flowering time. Nature Genetics 28: 286-289.

Virk, PS, Ford-Lloyd, BV, Jackson, MT, and Newbury, HJ. 1996. Predicting quantitative variation within rice germplasm using molecular markers. Heredity 76: 296-304.

Vos, P, Hogers, R, Bleeker, M, Reijans, M, van de Lee, T, Hornes, M, Frijters, A, Pot, J, Peleman, J, Kuiper, M, and Zabeau, M. 1995. AFLP: a new technique for DNA fingerprinting. Nucleic Acids Research 23: 4407-4414.

Yu, J and Buckler, ES. 2006. Genetic association mapping and genome organization of maize. Current Opinion in Biotechnology 17: 155-160.

Chapter 4

Mapping whole populations through linkage disequilibrium: Comparing measures of allelic association and validating in an existing dataset.

4.1. Introduction.

The hybrid-polyploid-aneuploid nature of the sugarcane genome and the reproductive biology of the sugarcane plant determine the types of breeding strategies possible for variety improvement programmes. Because of the polyploid genome, the creation of inbred lines is not feasible. The decay in heterozygosity per generation for ploidy of degree $2k$ is given by the recurrence relationship $H = H'(4k - 3)/(4k-2)$ (Li 1978), where H is the level of heterozygosity in generation g, and H' in generation $g-1$. It follows that in a diploid species, seven generations of selfing are sufficient to result in homozygosity of more than 99%, whereas an octaploid (k=4) or decaploid (k=5) would require 65 and 85 generations respectively to achieve the same degree of homozygosity. This length of time required makes it impractical to develop inbred and isogenic sugarcane lines. In addition, many sugarcane genotypes are male sterile, which makes it impossible to create inbred lines even if it were desirable. As a result, sugarcane breeding programmes generally follow some form of recurrent mass selection scheme (Jackson, 2005), using a large number of parent genotypes selected on multiple yield, sucrose, pest and disease criteria. In the SASRI breeding programme, approximately 250 parent genotypes are used in crossings each year in bi-parental or poly-cross combinations.

Within this large collection of parental germplasm, multiple alleles for loci controlling traits of interest will be present. Due to the polyploid nature of the sugarcane genome, multiple alleles for individual loci of interest may also be present within single genotypes. This complicates the use of markers for quantitative trait alleles (QTAs) in breeding, as different alleles at the same locus may interact with each other in determining trait phenotype, and dosage effects of QTAs could also be significant. For example, the phenotypic effect of a single copy of a QTA associated with disease resistance could be nullified by the presence of several copies of QTAs associated with susceptibility at the same locus. The background effect of multi-allelic variation at a single locus thus means that information on the presence or absence of individual markers or QTAs may not be informative, or useful in making breeding decisions. Interactions may also exist between different loci involved in the phenotypic expression of the same quantitative trait. In order to take different allelic effects at the same locus into account, as well as interactions between loci, a genetic map is

required that provides information about allelic variation across loci, and not just the presence or absence of individual, independent markers.

In Chapter 3 it was argued that markers identified in a single segregating cross are not particularly useful in applied breeding programmes, as they only represent a subset of the alleles present in the whole population. In a similar fashion, a genetic map of one or two parental genotypes constructed from a selfed or bi-parental population will not be useful in a breeding programme where large numbers of varieties are used in crossing. At the same time, it is not practical to map all parents of interest using conventional bi-parental or selfed mapping populations.

A potential solution to these problems is to use linkage disequilibrium present within the breeding population to identify haplotypes inherited without (or with limited) recombination from common *Saccharum* and early sugarcane ancestors. As described in Chapter 2, linkage disequilibrium is predicted to be extensive in sugarcane due to the small number of ancestral clones contributing to most germplasm (Arceneaux, 1965), and the limited number of generations completed since the original hybridization events. This was demonstrated by Jannoo *et al.* (1999), who found that linkage disequilibrium was common in regions of the sugarcane genome up to 10cM in distance, and could sometimes be found between markers up to 30cM apart. In addition, linkage disequilibrium can be maintained within populations due to selection (Hartl and Clark, 1989.). Thus within a breeding population selected for the same traits over generations, linkage disequilibrium may be maintained within those genic regions containing important quantitative trait loci. Groups of marker-pairs in linkage disequilibrium can be assembled into linkage groups using standard methods (see section 2.2.5). The map resulting from linkage group assembly of markers in disequilibrium with a population will represent haplotypes present in the most recent common ancestors (MRCA) contributing to the breeding population.

Many metrics have been proposed as measures of linkage disequilibrium, or non-independence of alleles at two loci (Morton *et al.*, 2001, Pritchard and Przeworski, 2001). These measures are sensitive to differences in marker allele frequency (Morton *et al.*, 2001), which is particularly relevant for sugarcane. Unlike diploids, in a polyploid a marker at low frequency may be linked to a marker at high frequency, which poses some challenges in terms of detecting disequilibrium. The objective of this chapter is to compare by simulation a common measure of disequilibrium – the correlation coefficient, r – with the association probability – ρ – described by Morton *et al.*, 2001 as an optimum measure of allelic association, and derive significance thresholds for use in mapping. The square of the

correlation coefficient, r^2, is generally considered a standard measure of linkage disequilibrium (McVean, 2002, Pritchard and Przeworski, 2001). In the context of whole-population mapping, I chose to use r instead, as it preserves information about the sign of the association – i.e. positive or negative. In genetic linkage analysis, a negative association would imply linkage in repulsion. Although linkage in repulsion due to preferential pairing of homologs is commonly detected in sugarcane mapping populations (Grivet *et al.*, 1996, Ming *et al*, 1998, Aitken *et al.,* 2005), one would not expect that this is conserved within a population of diverse germplasm. Any negative association detected would therefore be spurious. If r^2 were used at the measure of disequilibrium, the distinction between positive (true) and negative (false) association would be lost, resulting in an increase in the Type I error rate.

Following the comparison of the two measures of association, the more appropriate metric will be validated using an existing data set consisting of markers from a collection of germplasm that had been scored relative to a reference map of the sugarcane variety R570. This will indicate if a map of haplotypes derived from estimates of linkage disequilibrium within a diverse population has a physical interpretation in individual genotypes, and would be a useful tool in breeding.

4.2. Materials and Methods.

4.2.1. Linkage disequilibrium metrics.

Both r and ρ are derived from the covariance, D, between alleles present at two loci, A and B, where $D = |\pi_{AB}\pi_{00} - \pi_{A0}\pi_{0B}|$, and π is the frequency of each genotypic class, and A (or B) the presence of a marker at the locus, and 0 the absence. The correlation coefficient, r, is then $r = D/\sqrt{Q(1-Q)R(1-R)}$, and the association probability $\rho = D/Q(1-R)$, where Q and R are the frequencies of markers A or B with the condition that $Q \leq R, 1-Q$ (Morton *et al.*, 2001). It is immediately obvious that the difference in the two metrics lies in the denominator, specifically in terms of the use of the frequencies of A and B. In order to illustrate the effect this has, the two metrics were calculated for a set of simulated data assuming 80 genotypes, where Q < R, and where A is always completely associated with B.

4.2.2. Simulation of significance thresholds.

For an average data set comprising of between 1400 to 1500 markers, there are in the order of one million marker pairs for which association can be calculated. If the detection of 10 false associations out of 1×10^6 combinations is acceptable, this corresponds to a Type 1 error probability of 1×10^{-5}. The value of r with 79 degrees of freedom corresponding to this

error probability is 0.455. Probability tables are not available for ρ, so were derived by simulation. Marker data (1 or 0) was randomly simulated at two loci for intermediate marker frequency (~0.5), and used to calculate ρ. This was repeated 1 x 10^6 times, and the 10 results with the highest value of ρ were reported. This was repeated 10 times, and the average of each of the tenth-ranked values of ρ was used as the significance threshold for subsequent simulations. The simulation was performed in the GAUSS^{tm} 7.0 programming language. In order to check the validity of the simulation, the same was done for the r metric, in order to compare the simulated result to the tabular value.

4.2.3. Simulation of Type 1 error probability.

Following the establishment of the ρ threshold a similar simulation was performed, this time calculating both *r* and ρ for the same set of simulated data on 80 genotypes at a range of marker frequencies, with each marker frequency combination being repeated 1 x 10^6 times. The frequency of *r* and ρ exceeding their individual 1 x 10^{-5} error probabilities was calculated for each simulation, and used to compare the relative effectiveness of the two metrics. The simulation was performed in the GAUSS^{tm} 7.0, and the GAUSS code is given in Appendix A.

4.2.4. Validation in a sugarcane population.

Through research collaboration with CIRAD (Centre de Cooperation Internationale en Recherche Agronomique pour le Développement, Montpellier, France), data was available from a collection of 74 sugarcane genotypes derived from various breeding programmes throughout the world. The data consisted of 1626 AFLP markers derived from 42 primer pairs, as described by Raboin (2005). Marker scoring had been done using AFLP autoradiographs from the R570 mapping population as a reference, and 417 of the 1626 markers scored were located on the R570 map. Thus any linkages involving these markers found to be significant in the population of 74 genotypes could be verified on the existing genetic map.

Values of *r* and ρ were calculated for all 86 736 marker-pair combinations where both markers occurred on the R570 map, using an algorithm written in GAUSS 7.0. Associations where both markers were located in the same linkage group in R570 were counted as correct, and cases where markers belonged to different R570 Homology Groups were counted as incorrect. Cases were associations were between markers in the same R570 Homology Group but assigned to different linkage groups were not counted, as these could represent specific recombination events in R570. If this is true, the count of correct assignment may be biased downwards.

4.2.5. Map construction.

Significant marker-pair associations were considered to be physically linked on the same haplotype. Groups of pairs with individual markers in common were assembled into linkage groups using standard mapping methodology. Linkage disequilibrium metrics (i.e. r and ρ) were considered to be equivalent to 1 – recombination coefficient, c, and markers were ordered using a branch and bound algorithm to minimise the sum of adjacent recombination coefficients, *sar*. (Weir, 1996). For example, consider loci 1, 2, 3, 4,...X. Start with pair 1:2. Locus 3 can be added in 3 possible ways: 3:1:2; 1:3:2; 1:2:3. Calculate the *sar* for each combination, where *sar* for combination 3:1:2 = $c_{3:1} + c_{1:2}$. Find the combination with the smallest *sar*, and add locus 4 to that combination in all possible orders. The process is then repeated until all loci have been added. As there is no valid interpretation of r in terms of physical distance in centi-Morgans, markers were ordered, but not assigned relative positions or distance estimates. The algorithms to perform the ordering of markers in linkage groups were implemented within GAUSS 7.0.

4.3. Results.

4.3.1. Comparison of the metrics r versus ρ.

As described under 4.2.1 above, the difference between r and ρ lies in the term used as the divisor of the covariance between markers. Morton *et al.*, (2001) proposed ρ as the optimal measure of association as it is insensitive to frequency differences between markers. This is illustrated in Table 4.1.

Table 4.1. Change in value of r and ρ for markers at different frequency, when association is complete.

Q	R	r	ρ
0.10	0.10	1	1
0.10	0.15	0.79	1
0.10	0.20	0.67	1
0.10	0.25	0.58	1
0.10	0.30	0.51	1
0.10	0.35	0.45	1

Table 4.1 gives values for r and ρ from simulated data of 80 genotypes, where the frequencies, Q and R of markers A and B vary, but association between the less frequent marker Q and the more frequent marker R is always complete. It can be seen that as the frequency of marker B diverges from that of marker A, the correlation coefficient, r

decreases, whereas ρ remains constant, reflecting the complete association of A with B. In theory therefore, in cases where low frequency markers are physically linked to markers at high frequency, ρ should provide a better estimate of association than *r*.

4.3.2. Simulation of significance threshold for ρ and r.

Results of 10 repeat simulations of 1 x 10^6 random marker-pairs with Q and R = 0.5 gave the value of an average Type 1 error probability of 0.00001 as ρ = 0.61 (standard error = 0.016; range = 0.58 to 0.63). This value was used in subsequent simulations comparing *r* versus ρ. Results for *r* gave a value of *r* = 0.48 (standard error = 0.007; range = 0.47 to 0.49). This was of the same order as the tabular value of 0.45, but to be consistent the higher value from the simulation was used for subsequent tests.

4.3.3. Simulation of Type 1 error rate for r and ρ at different marker frequencies.

Simulation was done for different frequencies of markers A and B. For each frequency combination, five repeat simulations of 1 x 10^6 marker-pairs were done, the number of associations exceeding the thresholds set for *r* (0.48) and ρ (0.61) were counted and the mean and standard error calculated. Results are presented in Table 4.2.

For the case of Q = R = 0.5, the rate of detecting random associations as significant is 7 per 10^6 tests and 8 per 10^6 for *r* and ρ respectively, which is in line with the rate of 10 per million predicted from simulating the threshold. It is seen that for low values of Q and R, both metrics have high Type 1 error rates, with ρ being particularly prone to detecting random associations as significant. For most values of Q, as R increases, the error rate for *r* decreases, while that for ρ increases. This indicates that although ρ may be less sensitive to variation in marker frequency with regards to Type 2 errors (incorrectly rejecting the null hypothesis) as illustrated in 4.1, it is highly prone to Type 1 errors as marker frequencies diverge. It appears therefore, that *r* will be a more appropriate metric to use in estimating linkage disequilibrium in this context than ρ.

The correlation coefficient gave acceptable error rates from Q between 0.3 and 0.7, and at lower values of Q as R diverged in frequency. In order to obtain alternative thresholds for *r* for low marker frequencies, the simulation of significance thresholds for *r* was repeated as described in section 4.2.3 above, for values of Q = R = 0.2 and Q = R = 0.1, as well as Q = R = 0.8. The value of *r* for these frequencies for which the Type 1 error rate is 1 x 10^{-5} is given in Table 4.3. A similar correction was not tried for the ρ metric, as it gave unacceptably high Type 1 errors at all values of Q.

Table 4.2. Numbers of false associations out of 1 x 10^6 tests for r and ρ at thresholds of 0.48 and 0.61 respectively, for different marker frequencies, Q and R.

Q	R	r	ρ	s.e.(r)	s.e.(ρ)
0.1	0.1	607	2496	23	56
0.1	0.2	119	4209	6	79
0.1	0.3	28	12920	6	67
0.2	0.2	55	140	3	9
0.2	0.3	26	484	6	23
0.2	0.4	10	2034	3	17
0.3	0.3	15	26	3	5
0.3	0.4	12	129	6	4
0.3	0.5	3	764	2	34
0.4	0.4	10	11	5	3
0.4	0.5	8	75	2	12
0.4	0.6	4	579	1	27
0.5	0.5	7	8	1	3
0.5	0.6	6	73	3	5
0.5	0.7	6	762	3	16
0.6	0.6	9	9	4	2
0.6	0.7	9	122	3	10
0.6	0.8	10	2061	4	53
0.7	0.7	16	27	4	3
0.7	0.8	23	478	4	27
0.8	0.8	56	133	8	11

Table 4.3. Simulated value of r giving a Type 1 error probability of 1 x 10^{-5}

Q	R	r
0.1	0.1	0.67
0.2	0.2	0.53
0.8	0.8	0.54

4.3.4. Validation against existing map.

Values of r and ρ were calculated for all 86 736 pairs of markers for which map position in the variety R570 was available. Initial significance thresholds of 0.48 and 0.61 were used for r and ρ respectively, for comparative purposes. From the results in Table 4.3 above, a second threshold for r was used to see if Type I errors could be reduced, while maintaining a

low Type II error rate. This was done by making *r* conditional on Q by increasing the *r* threshold to 0.53 for associations where at least one marker had a frequency $0.2 < Q < 0.3$, and to 0.67 for $Q < 0.2$.

The numbers of marker-pair associations in a population of 74 diverse genotypes that were correctly (or incorrectly) assigned to known linkage groups based on the map of the variety R570 are shown in Table 4.4. Comparing 4a with 4c shows that although ρ again detects a greater number of real associations that *r*, it also has a very high rate of declaring false associations as significant, to the extent that at the threshold limit, it detects more false associations than real associations. A threshold value of 0.55 for *r* results in an acceptable number of Type 1 errors, but ρ has a very high number of Type 1 errors even at high threshold values.

Table 4.4. Associations correctly and incorrectly identified by *r* and ρ at different threshold values.

a.

	$r > 0.48$		
Threshold	# cases	# incorrect	# correct
0.80	53	0	53
0.75	70	0	70
0.70	80	0	80
0.65	95	0	95
0.60	122	0	118
0.55	159	3	149
0.50	232	17	182
0.48	247	18	192

b.

	***r* conditional on Q**		
Threshold	# cases	# incorrect	# correct
0.80	53	0	53
0.75	70	0	70
0.70	80	0	80
0.65	95	0	95
0.60	121	0	117
0.55	155	2	144
0.50	206	10	171
0.48	217	11	179

c.

	$\rho > 0.61$		
Threshold	# cases	# incorrect	# correct
0.80	255	39	135
0.75	386	88	162
0.70	543	138	200
0.65	749	217	233
0.61	932	286	259

d.

Associations with:	*r*	ρ
Missed by ρ	25	
Correctly rejected by ρ	28	
Missed by *r*		92
Correctly rejected by *r*		645

Although a threshold of 0.55 for *r* has a lower Type I error rate than a threshold of 0.48, increasing the threshold would result in 43 Type II errors – i.e. rejecting an association when it is known to be real. Using an *r* threshold conditional on frequency Q as described above results in small decrease in Type I errors, but a substantial increase in Type II errors (Table 4.4b).

In order to try and understand the discrepancies in Type I and II errors between r and ρ, the data was examined in more detail. Values of r and ρ for the same marker pairs were filtered to find cases where correct associations found by *r* were missed by *ρ*; incorrect associations found by *r* were correctly rejected by *ρ*, and visa versa. These results are shown in Table 4.4d. Again it is seen that in relative terms the number of Type II errors for *r* (92) is higher than that for *ρ* (25), while the number of Type I errors for *r* (28) is substantially lower than that for *ρ* (645). In terms of efficiency of detecting true associations, it is important to understand the conditions under which ρ detects true associations that are missed by *r*. Table 4.5 shows a subset of marker-pair associations correctly identified by *ρ*, but not declared significant by *r* at the threshold used (0.48).

Table 4.5. Associations correctly identified by ρ, but not identified by *r*.

Marker1	Marker2	*r*	*ρ*	Q	R	R/Q	LG1*	LG2*
64r11	23rm8	0.29	1	0.31	0.84	2.7	IV-2	IV-2
47r7	23rm8	0.34	1	0.36	0.83	2.3	IV-2	IV-2
23r13	68r5	0.35	1	0.26	0.74	2.9	VIII-2	VIII-2
44r8	23rm8	0.31	1	0.33	0.84	2.5	IV-2	IV-2
44r8	63r16	0.44	1	0.33	0.72	2.2	IV-2	IV-2
46r18	63r16	0.46	1	0.36	0.73	2.0	IV-2	IV-2
67r12	24rm18	0.39	1	0.40	0.81	2.0	VIII-d	VIII-d
74r12	46rm10	0.42	1	0.38	0.78	2.1	II-a	II-a
36rm28	58rm30	0.43	1	0.42	0.79	1.9	VI-1b	VI-1b
53rm8	22rm23	0.45	1	0.39	0.76	1.9	VI-11	VI-11

* LG for markers 1 and 2 refers to the homology goup and linkage group in the map of R570. e.g. IV-2 refers to linkage group 2 in homology group IV.

As can be seen from Table 4.5, the value of *ρ* indicates a complete association for a marker at low frequency, with a marker at high frequency, illustrating the insensitivity of *ρ* to marker frequency in terms of Type II errors. These cases were not detected as significant by *r*. The genetic interpretation and use of these associations in breeding is difficult, however (see section 4.4 below). It therefore appears that although *r* has a relatively high level of Type II errors compared to ρ, the specific cases involve associations that are relatively non-informative, or difficult to use in a practical manner. In light of this, further map construction was done using *r* as the more appropriate measure of association in this context.

4.3.5. Map construction.

The mapping algorithm was then performed on the list of significant marker-pair associations as summarised in Tables 4.4a and 4.4b. Using all marker pairs for which $r > 0.48$ resulted in a map M1 of 180 markers assigned to 54 linkage groups, and using r conditional on Q gave a map M2 of 166 markers in 51 linkage groups (LGs). In M1, six LGs contained markers from different Homology Groups in the R570 map, caused by false associations. In M2, three LGs contained incorrectly assigned markers. As the two maps were very similar and M2 was not superior to M1, the map M1 was used for further investigation.

The TropGENE database (http://tropgenedb.cirad.fr) was interrogated for map information for the variety R570. Data on the map position in centi-Morgans for the AFLP markers used for the validation set was downloaded for two independent mapping populations used to derive the R570 map. The first data set was that of Rossi *et al.* (2003), which had been used to assign linkage groups to Homology Groups. This map was made from a population of selfed progeny of R570. The second data set was that of Raboin *et al.* (2006), constructed from the bi-parental cross R570 x MQ76/53. Map position was available for all markers from the Rossi map, but only for some markers from the Raboin map. The distance in centi-Morgans for each group of markers in linkage disequilibrium within the validation population of 74 genotypes was calculated from the marker position in the Rossi map, provided they were from the same Homology Group. Table 4.6 shows the linkage groups in disequilibrium in the validation population with their corresponding Homology Groups and linkage groups in the R570 map, the map positions in the Rossi and Raboin maps with the Raboin linkage groups, and the extent of linkage disequilibrium in centi-Morgans. Markers highlighted in bold itallics are false linkages. Markers belonging to un-assigned linkage groups on the R570 map may be false associations, or may be real associations not detected in R570.

False linkages and unassigned markers were not included in calculating the extent of linkage disequilibrium. Table 4.7 summarises the extent of linkage disequilibrium in the LGs detected, and shows that although most haplotypes in disequilibrium are less than 10 cM in size, several regions exists where LD extends more than 50 cM. The largest linkage group detected was 120 cM in length.

Comparing markers in the LD linkage groups with map positions (Table 4.6) shows that the relative order of markers generally differs between the LD and bi-parental maps, but also differs between the two bi-parental maps. For example for R570 HG/LG 1-a, the two markers furthest apart in the Rossi reference map – aagcag6 and acactg18 – are located in the middle of the LD derived linkage group 2. These markers are also located in the middle

Table 4.6. Comparison of linkage groups derived through disequilibrium with two independently derived maps of variety R570. The assignment of homology groups and linkage groups (HG/LG) is per the Rossi *et al.* (2003) map. Markers where the homology group is 'U' were not assigned to any HG in the Rossi map.

Marker code	Marker name	LD linkage group	R570 HG/LG	Rossi map cM	cM LD Rossi map	Raboin map cM	Raboin map LG
12r28	aaccac28	2	I-a	15.90		79.3	LG54
64r10	actcat10	2	I-a	14.00		80.2	LG54
23r6	aagcag6	2	I-a	12.10		79.3	LG54
37r18	acactg18	2	I-a	25.80		71.4	LG54
36r31	acactc31	2	I-a	23.70	13.70	70.0	LG54
37rm25	acactg25	30	I-c	19.10		*	*
71r28	agccaa28	30	I-c	0.00	19.10	*	*
47r9	accctg9	38	I-4	37.60		54.1	LG43
62rm14	actcac14	38	I-8	33.40	4.20	*	*
12r10	aaccac10	44	I-4	6.70		68.8	LG43
37r15	acactg15	44	I-4	4.80	1.90	73.2	LG43
53r5	***acgcag5***	***29***	***VI-2***	***227.90***		***0.0***	***LG173***
46rm11	accctc11	29	I-5	54.10		*	*
47r8	accctg8	29	I-5	46.00	8.10	32.8	LG154
47r18	accctg18	6	I-7	41.60		81.8	LG40
26r11	aagctc11	6	I-7	50.40		86.6	LG40
76r7	agcctc7	6	I-7	51.60	10.00	90.4	LG40
74r28	agccat28	7	I-7	164.90		40.7	LG40
68r13	actctt13	7	I-7	195.50		33.1	LG40
46r6	accctc6	7	I-7	189.10		*	*
44r5	acccat5	7	I-7	189.10	30.60	36.3	LG40
36r27	***acactc27***	***37***	***IV-1***	***3.70***		*	*
76r14	agcctc14	37	U-11	81.20		7.1	LG164
35rmF	acactaF	37	I-7	221.70		*	*
31r7	acacaa7	37	I-7	226.90		15.3	LG40
55r24	acgcta24	37	I-7	232.10	10.40	0.0	LG40
84r19	***aggcat19***	***37***	***II-9***	***54.00***		***27.2***	***LG69***
44r23	acccat23	23	I-9	34.40		29.4	LG108
67r3	actctg3	23	I-9	35.40	1.00	28.9	LG108
71r1	agccaa1	36	II-a	92.50		22.2	LG128
58r10	acgctt10	36	II-a	82.10	10.40	11.5	LG128
46r5	accctc5	46	II-a	66.40		*	*
86r5	aggctc5	46	II-a	63.30	3.10	3.9	LG128
58r6	acgctt6	5	II-b	47.00		23.1	LG166
46r12	accctc12	5	II-b	47.80		22.5	LG166
22r18	aagcac18	5	II-b	90.40		0.0	LG166
21r10	aagcaa10	5	II-b	48.60	43.40	*	*

Table 4.6 continued.

Marker code	Marker name	LD linkage group	R570 HG/LG	Rossi map cM	cM LD Rossi map	Raboin map cM	Raboin map LG
53rB	acgcagB	22	II-b	0.00		*	*
42r17	acccac17	22	U-27	73.00		29.3	LG176
24r1	aagcat1	22	U-27	56.00		*	*
22r15	aagcac15	22	U-27	25.60	47.40	0.0	LG176
84r24	aggcat24	39	II-6	82.20		8.9	LG21
44r12	acccat12	39	II-6	83.20	1.00	12.1	LG21
58rm8	acgctt8	12	II-7	21.20		*	*
21rm9	aagcaa9	12	II-7	21.20	0.00	*	*
44r33	***acccat33***	***26***	***VIII-11***	***36.70***		***6.7***	***LG56***
55rm17	acgcta17	26	II-9	16.40		*	*
26rm4	aagctc4	26	II-9	33.50		*	*
87r9	aggctg9	26	II-9	38.30	21.90	9.8	LG69
86r12	aggctc12	51	III-3	43.90		48.0	LG53
73rm15	agccag15	51	III-3	44.00	0.10	*	*
44rm37	acccat37	53	III-3	98.70		*	*
68rm10	actctt10	53	III-3	102.00	3.30	*	*
12rm7	aaccac7	43	III-4	127.30		*	*
31rm5	acacaa5	43	III-4	127.10	0.20	*	*
25r26	aagcta26	45	U-31	0.00		*	*
28r40	aagctt40	45	III-8	65.60		*	*
36r13	acactc13	45	III-8	42.80	22.80	26.3	LG27
53r23	acgcag23	50	U-30	0.00		4.9	LG171
44r22	acccat22	50	III-10	12.00		0.0	LG57
84r20	aggcat20	21	IV-1	149.30		12.6	LG81
23r28	aagcag28	21	IV-1	155.10		6.2	LG81
87rm14	aggctg14	21	IV-1	156.00		*	*
63r6	***actcag6***	***21***	***VIII-11***	***28.80***		***9.8***	***LG56***
71r7	***agccaa7***	***21***	***U-3***	***14.00***		***41.4***	***LG111***
64r24	***actcat24***	***21***	***U-49***	***5.00***		*	*
78r13	agcctt13	21	IV-1	137.90	18.10	*	*
44r11	***acccat11***	***21***	***U-21***	***7.00***		***2.3***	***LG107***
47rR999	accctgR	3	U-52	109.70		*	*
67r13	actctg13	3	U-52	116.70		9.3	LG12
37r2	acactg2	3	IV-2	91.00		37.6	LG106
46r18	accctc18	3	IV-2	104.60		28.4	LG106
44r9	acccat9	3	IV-2	161.40		33.6	LG106
47r7	accctg7	3	IV-2	126.40		19.8	LG106
44r8	acccat8	3	IV-2	116.30		21.9	LG106
64r11	actcat11	3	IV-2	118.60		17.1	LG106
76rm25	agcctc25	3	IV-2	122.50	70.40	*	*

Table 4.6 continued.

Marker code	Marker name	LD linkage group	R570 HG/LG	Rossi map cM	cM LD Rossi map	Raboin map cM	Raboin map LG
55r11	acgcta11	24	IV-2	66.20		67.1	LG106
37r32	acactg32	24	IV-2	64.60		68.2	LG106
36r29	acactc29	24	IV-2	85.70		42.9	LG106
85r14	aggcta14	24	IV-2	90.30		*	*
67r10	actctg10	24	IV-2	87.20	25.70	40.8	LG106
44rm24	acccat24	33	IV-2	11.40		*	*
37rm9	acactg9	33	IV-2	5.30	6.10	*	*
23rm2	aagcag2	20	VI-c	107.60		*	*
64r31	actcat31	20	VI-c	101.80	5.80	15.7	LG11
28r2	aagctt2	47	VI-c	58.30		36.7	LG11
74r7	agccat7	47	VI-c	56.50	1.80	37.9	LG11
62rm21	actcac21	27	VI-d	44.10		*	*
86r24	aggctc24	27	VI-d	31.90		23.8	LG65
85r2	aggcta2	27	VI-d	5.00		10.1	LG65
55r18	acgcta18	27	VI-d	5.90	39.10	9.6	LG65
44rm10	acccat10	10	VI-1a	77.40		*	*
55rm3	acgcta3	10	VI-1a	77.40		*	*
21rm24	aagcaa24	10	VI-1a	106.70		*	*
26rm16	aagctc16	10	VI-1a	134.00	56.60	*	*
64r6	actcat6	49	VI-1a	157.20		0.0	LG179
53r16	acgcag16	49	U-4	47.80		11.1	LG172
42r15	***acccac15***	***14***	***III-4***	***117.50***		***0.0***	***LG191***
68r14	***actctt14***	***14***	***IV-2***	***85.70***		***41.3***	***LG106***
42rm10	acccac10	14	VI-1b	63.40		*	*
37rm23	acactg23	14	VI-1b	61.50		*	*
36rm28	acactc28	14	VI-1b	66.20	4.70	*	*
21r1	aagcaa1	41	VI-2	66.10		*	*
25r37	aagcta37	41	VI-2	135.60	69.50	*	*
21rm20	aagcaa20	4	VI-3	159.10		*	*
44r16	acccat16	4	VI-3	93.40		125.6	LG7
12r13	aaccac13	4	VI-3	61.10		142.8	LG7
47r14	accctg14	4	VI-3	86.30		128.7	LG7
41r14	acccaa14	4	VI-3	99.60		121.9	LG7
23r14	aagcag14	4	VI-3	39.40		*	*
58rm20	acgctt20	4	VI-3	44.40	119.70	*	*
21r19	aagcaa19	9	VI-3	209.30		*	*
67r6	actctg6	9	VI-3	209.30		0.0	LG59
23r10	aagcag10	9	VI-3	211.20		5.6	LG59
12rm32	aaccac32	9	VI-3	217.60	8.30	*	*

Table 4.6 continued.

Marker code	Marker name	LD linkage group	R570 HG/LG	Rossi map cM	cM LD Rossi map	Raboin map cM	Raboin map LG
28rm25	aagctt25	16	VI-3	254.80		*	*
12rm30	aaccac30	16	VI-3	254.80		*	*
31r20	acacaa20	16	VI-3	246.70		64.9	LG59
64r22	actcat22	16	VI-3	254.80	8.10	62.8	LG59
84r10	aggcat10	42	VI-4	36.50		17.3	LG125
78r29	agcctt29	42	VI-4	38.40	1.90	19.4	LG125
62rm26	actcac26	42	U-11	18.10		*	*
26rI	aagctcI	35	VI-10	7.60		21.1	LG66
38r22	acactt22	35	VI-10	13.90	6.30	23.2	LG66
53rm8	acgcag8	17	VI-11	0.00		*	*
12rm16	aaccac16	17	VI-11	36.10	36.10	*	*
48r14	accctt14	13	VI-12	36.80		7.8	LG113
87rA	aggctgA	13	VI-12	83.90		4.6	LG113
53r22	acgcag22	13	VI-12	6.00		4.1	LG113
35r10	acacta10	13	VI-12	25.10	77.90	0.0	LG113
62rm25	actcac25	31	VII-8	15.20		*	*
21rm15	aagcaa15	31	VII-8	15.20	0.00	*	*
38rm19	acactt19	1	VIII-a	17.40		*	*
63rm10	actcag10	1	VIII-a	16.40		*	*
26rm5	aagctc5	1	VIII-a	0.00		*	*
46rm19	accctc19	1	VIII-a	15.40		*	*
46r21	accctc21	1	VIII-a	51.10	51.10	19.1	LG184
71rA	agccaaA	11	VIII-e	41.20		0.0	LG71
26r8	aagctc8	11	VIII-e	46.10	4.90	3.0	LG71
35r15	acacta15	8	U-61	29.60		0.0	LG115
63r23	actcag23	8	VIII-1	34.50		102.1	LG67
87r13	aggctg13	8	VIII-1	36.30		90.7	LG67
26rH	aagctcH	8	VIII-1	65.60	31.10	83.1	LG67
23r33	aagcag33	40	VIII-1	156.90		39.7	LG67
85rmF	aggctaF	40	VIII-1	180.30		*	*
26r12	aagctc12	40	VIII-1	182.30		9.7	LG67
44r27	acccat27	40	VIII-1	180.30		*	*
71r24	agccaa24	40	VIII-1	215.00	58.10	0.0	LG67
73r2	agccag2	25	VIII-2	14.10		108.5	LG5
23r13	aagcag13	25	VIII-2	43.00		*	*
86r21	aggctc21	25	VIII-2	15.00		106.4	LG5
75rm4	agccta4	25	VIII-2	41.90	28.90	*	*

Table 4.6 continued.

Marker code	Marker name	LD linkage group	R570 HG/LG	Rossi map cM	cM LD Rossi map	Raboin map cM	Raboin map LG
64rm5	actcat5	18	VIII-2	127.50		*	*
62r22	actcac22	18	VIII-2	139.60		49.0	LG5
36r15	acactc15	18	VIII-2	99.20		60.8	LG5
42rm25	acccac25	18	VIII-2	108.80		*	*
26r2D	aagctc2D	18	U-52	18.20		*	*
44r39	acccat39	18	VIII-2	182.10	82.90	20.6	LG5
12r4	***aaccac4***	***48***	***VI-c***	***63.90***		***34.1***	***LG11***
24r4	aagcat4	48	VIII-11	28.60		*	*
31r13	acacaa13	48	VIII-11	42.10	13.50	7.2	LG56
31r6	acacaa6	52	U-6	31.00		*	*
67rJ	actctgJ	52	U-6	31.60	0.60	40.1	LG137
67rG	actctgG	34	U-8	29.60		9.3	LG114
85r8	aggcta8	34	U-8	39.70		19.1	LG114
42r16	acccac16	34	U-8	47.80	18.20	43.3	LG114
86r8	aggctc8	19	U-17	0.00		0.0	LG31
85r11	aggcta11	19	U-17	7.00		6.4	LG31
58r19	acgctt19	19	U-17	11.20	11.20	13.9	LG31
36r8	acactc8	54	U-33	0.00		*	*
12r9	aaccac9	54	U-25	13.90		0.0	LG145
37rm26	acactg26	15	U-44	0.00		*	*
64rm19	actcat19	15	U-44	2.70	2.70	*	*
28r20	aagctt20	28	U-52	21.50		42.5	LG12
37r7	acactg7	28	U-52	86.90		26.8	LG12
46rm23	accctc23	28	U-52	53.30		*	*
25r40	aagcta40	28	U-52	94.10	72.60	*	*
76rm15	agcctc15	32	U-36	15.90		*	*
26rmB	aagctcB	32	U-36	17.80	1.90	*	*

Table 4.7. Summary of the distribution of the extent of linkage disequilibrium detected among 54 haplotypes.

Extent of linkage disequilibrium in cM	Number of LGs
0-10	23
10-20	9
20-30	4
30-40	4
40-50	2
> 50	9

of the LG54 derived from the Raboin map of R570. In some cases, agreement between all three maps does exist, for example the fragment of R570 LG I-7 equivalent to LG 6 in the linkage disequilibrium map and LG 40 from the Raboin map.

Cases of separate linkage disequilibrium LGs which corresponded to the same LG on the R570 map allowed the comparison of the extent of linkage disequilibrium with linkage equilibrium. For example, LD LGs 6, 7 and 37 correspond to the same R570 LG I-7, which is 232.2 cM long on the Rossi map. Within the three LD segments disequilibrium extends for 10 cM, 30 cM and 10.4 cM respectively, representing fairly short sections of this chromosome. The distance between these segments is 113.0 cM (LD LG 6 vs 7) and 26.2 cM (LD LG 7 and 37). LD LG 37 contains the marker at the end of the Rossi map (acgcta24; 232.2 cM), so this section of the chromosome is present in the LD map. The beginning section however, is not present in the LD map, representing a section of 41.6 cM. The LD map covers a total of 50.4 cM, leaving 181.8cM unmapped by linkage disequilibrium. On the other hand, LD LGs 3, 24 and 33 correspond to R570 LG IV-2, which has a total length of 161.8cM. Together these three linkage disequilibrium LGs cover 102.2 cM, leaving only 59.2 cM unmapped. Although LD LGs 3 and 24 are themselves fairly long (70.4 cM and 25.7 cM respectively), the two fragments in LD are separated by only 0.7 cM on the R570 map. This may represent a case of a specific recombination event in the genotype of R570, compared to the sugarcane population at large.

Haplotypes were identified for different linkage groups within the same R570 Homology Groups. Assuming that the chromosomes/linkage groups in the R570 map are orientated in the same direction, markers in a similar map position on different LGs within a Homology Group may reflect allelic variation at loci within the region. For example, Homology Group VI is represented by linkage disequilibrium fragments on the four linkage groups c, 1a, 2 and 3 in the region of map position ~100 cM. If this region contained a major gene or QTL, examination of the association and size of the effect of the individual markers with the phenotype would reveal possible allelic variants or QTAs for the locus. This information would be of particular interest for marker assisted breeding in a polyploid crop, as it may allow the effect of allelic interactions on phenotype to be explained.

4.4. Discussion.

In this study, the hypothesis that linkage disequilibrium can be used to create a map depicting haplotypes within a population of sugarcane genotypes has been tested and validated. Two metrics, the correlation coefficient, r, and the association probability, ρ, were

compared by simulation to assess their relative efficiency in detecting real association between markers while minimising the spurious detection of random associations. The correlation coefficient was used instead of r^2, in order to preserve information about the sign of the association, and to reduce the Type I error rate, as described in section 4.1.

Comparison of the two metrics both by simulation and in the validation population showed that the use of the correlation coefficient resulted in the detection of far fewer random associations (Type I errors) than the association probability. This was because the association probability is insensitive to differences in marker frequency. The insensitivity of ρ to frequency differences is desirable if physical linkage within a candidate region is already known from an existing map, and the objective is to locate the position of a gene affecting a phenotype of interest within that region (Morton et al., 2001). Under these circumstances, the probability of Type I errors is not important. In terms of detecting unknown linkage however, this study showed that ρ is not suitable due to the high probability of declaring random associations as significant as marker frequencies diverge. Ironically, this same insensitivity to marker frequency differences resulted in ρ correctly identifying more real associations in the validation population (i.e. fewer Type II errors), as shown in Tables 4.4 and 4.5. These all involved cases where a marker at low frequency was linked to a marker at high frequency. At first sight, it may appear desirable to try and retain these associations, as they will increase the size of the subsequently derived map. When attempting to interpret and apply them in breeding, however, these associations may be problematic. In a polyploid, a marker at high frequency in the population is likely to be present in several copies within individual genotypes – i.e. present on several chromosomes within the same Homology Group. If marker A is present at high frequency (i.e. duplicated across homologs), and is associated with markers B and C (both at low frequency), then B and C will map to the same linkage group as A, based on their common association with A. The markers B and C may, however, be on different haplotypes within the same Homology Group, and are therefore not physically linked. This is described in more detail in Chapter 5.3.3 and Figure 5.4. Retaining true associations between markers where the frequency difference is large may thus result in assembling false physical linkage groups during mapping. In addition, high frequency markers are often not useful in terms of application in marker-assisted selection, as they are present in the majority of the population. In this case selection efficiency is often not increased by using marker information, as a selection of random germplasm would have a high probability of containing the marker. An exception to this, however, is the case of a marker linked to a negative allele – e.g. disease susceptibility. Under these circumstances, the marker could be used effectively to identify the rare cases of genotypes lacking the allele for susceptibility. Specific cases of mapping individual known marker-trait associations of

importance could be analysed separately, however, if it was thought necessary. For these reasons, it was decided that trying to retain associations between markers at different frequencies would not be appropriate for coarse-scale mapping of linkage disequilibrium within breeding germplasm, and the correlation coefficient was used for further mapping work.

Comparing the linkage disequilibrium map derived from a collection of diverse germplasm against an existing map of the sugarcane variety R570 showed that the LD haplotypes did have a valid interpretation in terms of physical linkage. In addition, comparison of the LD map against the genetic linkage map of R570 provided new information on the extent of linkage disequilibrium in sugarcane. More than half the LD linkage groups corresponded to chromosomal regions of more than 10 cM in size, with nine LGs corresponding to more than 50 cM. The largest LD linkage group corresponded to ~120 cM on the genetic map. The large variation in the extent of LD detected is likely due to the hybrid-aneuploid nature of the sugarcane genome. Jannoo *et al.* (1999) reported that two thirds of the marker-pairs in disequilibrium within a population of diverse sugarcane germplasm were derived from *S. spontaneum*, despite that fact that only 20% of the genome is of *S. spontaneum* origin. Some chromosomes of *S. spontaneum* origin may have limited pairing affinity with their *S. officinarum* homo(eo)logues, resulting in reduced recombination and more extensive linkage disequilibrium. Likewise, recombinant chromosomes containing segments derived from both *S. officinarum* and *S. spontaneum* may show reduced pairing (Jannoo *et al.* 1999), and tend to be transmitted intact through subsequent generations. If this is the case, whole-population LD maps in sugarcane may be over-represented by sections of the genome derived from S. spontaneum and by regions of recombination between *S. spontaneum* and *S. officinarum*.

The extent of linkage disequilibrium detected within this sugarcane population is considerably greater than that reported in other species to date. In *Arabidopsis*, Hagenblad and Nordborg (2002) found that LD extended up to 250 kb, equivalent to about 1 cM of recombination, in 14 sequenced fragments in the region of the *FRIGIDA* flowering locus. In maize Rafalski (2002) and Ching *et al.* (2002) reported that LD extends over 100 kb for the *adh1* and *y1* loci, and could not detect decay in LD over a 300-500 bp range for 18 other genes. Although their results are presented in the unit of base-pairs and not centiMorgans, the physical distance is less than that reported for *Arabidopsis*. In sorghum, Hamblin *et al.* (2004) estimated that linkage disequilibrium was sevenfold greater than that reported in maize, and suggested that LD was unlikely to extend further than 10 kb. Caldwell *et al.* (2006) detected LD up to 212 kb in elite barley population, similar to that found in *Arabidopsis*. Also in barley, Kraakman *et al.*

(2004) found the linkage disequilibrium was common in elite germplasm for markers within 10 cM distance on the integrated map.

The correlation between distance in base-pairs and distance in centiMorgans may vary within and between species due to variable recombination rates. Nevertheless, the extent of LD detected in section 4.3.5 is considerable larger than reported for *Arabidipsis*, maize, sorghum and barley. This confirms the presumption that LD is extensive in sugarcane due to its breeding history, as described in Chapter 2.7 and section 4.1 above.

Although the order of markers was not identical with the reference map used (Rossi *et al.* 2003; TropGENE database), the order of the Rossi map was also not identical with that of a second R570 map derived from a bi-parental cross (Raboin *et al.* 2006; TropGENE database). Exact marker order will depend on factors such as the size of the mapping population and the number of markers available, as well as possible genotyping errors. These errors could include cases where different DNA fragments are of the same size, and are scored as the same polymorphism. In addition, phenotypic selection within the mapping population of diverse germplasm may result in skewed estimates of recombination. For the level of resolution needed for coarse mapping genetic regions in linkage disequilibrium at the population level, the consensus between the LD map and the two bi-parental maps appears to be sufficient. The potential to reveal allelic variation across polyploid loci was also demonstrated by the identification of haplotypes at similar map positions in different linkage groups within the same Homology Group. This creates exciting new prospects in the molecular breeding of sugarcane as well as other polyploid crops, in terms of accounting for allelic variation and allelic interactions in the contribution of QTAs to phenotype.

4.5. References.

Aitken, KS, Jackson, PA and McIntyre, CL. 2005. A combination of AFLP and SSR markers provides extensive map coverage and identification of homo(eo)logous linkage groups in a sugarcane cultivar. Theoretical and Applied Genetics 110: 789-801.

Arceneaux, G., 1965. Cultivated sugarcanes of the world and their botanical derivation. Proceedings of the International Society of Sugar Cane Technologists 12: 844-854.

Caldwell, KS, Russell, J, Langridge, P and Powell, W. 2006. Extreme population-dependent diseuilibrium detected in an inbreeding plant species, *Hordeum vulgare*. Genetics 172: 557-567.

Ching, A, Caldwell, KS, Jung, M, Dolan, M, Smith, OS, Tingey, S, Morgante, M and Rafalski, AJ. 2002. SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BioMed Central Genetics 3: 19-32.

GAUSS Mathematical and Statistical System version 5. Aptech Systems Inc., Maple Vally, WA, 1984-2003.

Grivet, L, D'Hont, A, Roques, D, Feldmann, P, Lanaud, C and Glaszmann, JC. 1996. RFLP mapping in cultivated sugarcane (*Saccharum* spp): Genome organization in a highly polyploid and aneuploid interspecific hybrid. Genetics 142: 987-1000.

Hagenblad, J and Nordborg, M. 2002. Sequence variation and haplotype structure surrounding the flowering time locus *FRI* in *Arabidopsis thaliana*. Genetics 161: 289-298.

Hamblin, MT, Mitchell, SE, White, GM, Gallego, J, Kukatla, R, Wing, RA, Paterson, AH and Kresovich, S. 2004. Comparative population genetics of the Panicoid grasses: sequence polymorphism, linkage disequilibrium and selection in a diverse sample of *Sorghum bicolor*. Genetics 167: 471-483.

Hartl, DL and Clark, AG. 1989. Principles of Population Genetics, second edition. Sinauer Associates Inc, Sunderland, Massachusetts. pp 682.

Jackson, P.A. 2005. Breeding for improved sugar content in sugarcane. Field Crops Research 92: 277-290.

Jannoo, N, Grivet, L, Dookun, A, D'Hont, A and Glaszmann, JC. 1999. Linkage disequilibrium among sugarcane cultivars. Theoretical and Applied Genetics 99: 1053-1060.

Kraakman, ATW, Niks, RE, Van den Berg, PMMM, Stam, P and Van Eeuwijk, FA. 2004. Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars. Genetics 168: 435-446.

Li, CC. 1978. First Course in Population Genetics, second printing. The Boxwood Press, Pacific Grove, California. pp 631.

McVean, GAT. 2001. A genealogical interpretation of linkage disequilibrium. Genetics 162:987-991.

Ming, R, Liu, SC, Lin, YR, da Silva, J, Wilson, W, Braga, D, van Deynze, A, Wenslaff, TF, Wu, KK, Moore, PH, Burnquist, W, Sorrells, ME, Irvine, JE and Paterson, AH. 1998. Detailed alignment of *Saccharum* and *Sorghum* chromosomes: Comparative organization of closely related diploid and polyploid genomes. Genetics 150: 1663-1682.

Morton, NE, Zhang, W, Taillon-Miller, P, Ennis, S, Kwok, PY and Collins, A. 2001. The optimal measure of allelic association. Proceedings of the National Academy of Sciences USA 98: 5217-5221.

Pritchard, JK and Przeworski, M. 2001. Linkage disequilibrium in humans: Models and Data. American Journal of Human Genetics 69: 1-14.

Raboin, LM. 2005. Génétique de la résistance au charbon de la canne à sucre causé par *Ustilago scitaminea* syd.: caractérisation de la diversité génétique du pathogène, cartographie de QTL dans un croisement bi-parental et étude d'associations dans une population de cultivars modernes. Unpublished PhD thesis, Ecole Nationale Superieure Agronomique de Montpellier, France.

Raboin, LM, Oliveira, KM, Lecunff, L, Telismart, H, Roques, D, Butterfield, M, Hoarau, JY and D'Hont, A. 2006. Genetic mapping in sugarcane, a high polyploid, using bi-parental progeny; identification of a gene controlling stalk colour and a new rust resistance gene. Theoretical and Applied Genetics 112: 1382-1391.

Rafalski, A.. 2002. Applications of single nucleotide polymorphisms in crop genetics. Current Opinion in Plant Biology 5: 94-100.

Rossi, M, Araujo, PG, Paulet, F, Garsmeur, O, Dias, VM, Chen, H, Van Sluys, MA. and D'Hont, A. 2003. Genomic distribution and characterization of EST-derived resistance gene analogs (RGAs) in sugarcane. Molecular Genetics and Genomics 269: 406-419.

TropGENE database. http://tropgenedb.cirad.fr

Weir, BS. 1996. Genetic Data Analysis II. Sinauer Associates Inc, Sunderland, Massachusetts. pp 445.

CHAPTER 5

Population-level linkage disequilibrium mapping of haplotypes occurring within sugarcane breeding germplasm.

5.1. Introduction

As discussed in Chapters 3 and 4, molecular breeding in a polyploid crop may be confounded by allelic variation at quantitative trait loci (QTLs) involved in the phenotype of interest, as well as interaction of alleles (QTAs) within and/or between loci. The effectiveness of molecular breeding may be improved by the use of maps showing the haplotype composition for genetic regions of interest, as this would allow such interactions to be accounted for in a targeted manner. In the preceding Chapter, the methodology for constructing a map at the population-level using linkage disequilibrium was described.

Linkage disequilibrium mapping has been pioneered within the scientific community working on the human genome, with particular emphasis on mapping genes that cause disease (Reich *et al.*, 2001, Goldstein and Weale, 2001). The general approach that has been followed is to use LD within populations or pedigrees for the fine mapping of single nucleotide polymorphisms (SNPs) within a previously identified genetic region of interest known from an existing genetic map. Zhang *et al.* (2002) describe an LD map as being "*...constructed from a physical map with additive units for use in positional cloning by enhancing the resolution of the linkage map, for identifying sequences predisposing to recombination, and for discriminating other processes and events in population history.*" The scale of accuracy of mapping disease-causing genes with this method depends on the extent of linkage disequilibrium within the population studies. In a large-scale study, Reich *et al.* (2001) compared LD between a United States population of north European descent with that of a Nigerian population, and showed that LD typically extends for regions of 60 kb within the US population, but was far less within the Nigerian population. They conclude that genome-wide LD mapping is likely to be possible for populations with extensive LD, but that resolution may be limited to blocks in the range of 100 kb. For populations where LD blocks are small, fine structure mapping to identify specific disease-causing SNPs may be possible.

The situation with LD mapping in plants, however, is quite different, as the large resources of sequence and SNP information available for the human genome are generally not present. Although an appreciable number of studies have appeared in the literature with regards to using linkage disequilibrium to detect marker-trait associations (see Chapters 2 and 3), far

fewer reports of LD mapping in plants have been published. Those that have also rely on existing maps to interpret the associations in linkage disequilibrium detected, and do not address mapping *per se*. Kraakman *et al.* (2004, 2006) describe LD mapping of a number of yield, morphological and resistance traits in spring barley, compared to three existing barley maps. In a population of 146 barley cultivars, association was found between markers up to 10 cM apart. Many of the marker trait associations identified in the population were located in map regions where QTLs had been previously identified, but new putative QTLs were also identified, such as a new gene for resistance to barley yellow dwarf virus. Breseghello and Sorrels (2005), examined linkage disequilibrium in 95 winter wheat cultivars while accounting for the effects of within-population structure. On one chromosomal region studied, the extent of LD was less than 1 cM and could not be consistently detected at the marker density of the reference map. In a second centromeric region of chromosome 5A LD extended in the order of 5 cM. Previous evidence had suggested the presence of QTLs for kernel size in the chromosome regions studied, and this was confirmed by the identification of several significant associations between markers and kernel traits within the population. In potato, Simko *et al.* (2004) mapped an allele for verticillium wilt resistance in a diploid F1 population, and then tested its association with the disease within a collection of tetraploid potato germplasm. In a population of 137 individuals, the allele explained 21% of the variation in resistance. A recent review by Yu and Buckler (2006) discusses the opportunities of high-resolution QTL mapping in maize using linkage disequilibrium. The paradigm advocated in this review follows that used in human gene mapping in that it exploits new resources of DNA sequence information and platforms for large-scale genotyping for the fine mapping of candidate genes. The authors describe how association mapping has been useful in dissecting Mendelian traits, and conclude that over time LD mapping will address complex issues such as the understanding of dominance, epistasis, heterosis and genotype by environment interactions.

The studies summarised above have used existing genetic maps to provide a linkage disequilibrium interpretation to marker-trait associations detected within a population of cultivars, but have not used LD to map directly. In contrast, the objective of the work described in this Chapter is to construct a genetic map at the level of a breeding population by exploiting the unique breeding history of sugarcane that has resulted in fairly extensive population-level linkage disequilibrium. This has not been reported for any other crop species to date.

Linkage disequilibrium can result from a variety of causes, including physical linkage, genetic drift, population structure, admixture or growth, natural selection, variable recombination and

mutation rates and gene conversion (Palmer and Cardon, 2005). Within the context of physical linkage, the interpretation of a population-level map derived from estimates of linkage disequilibrium can be interpreted within the framework of the standard coalescent. The concept of the coalescent was first described by Kingman (1982a, 1982b, reviewed in Kingman, 2000.), in order to describe genetic processes during evolution. It is based on the idea of tracing the ancestry of a gene backwards in time, where in each generation a gene copy can be regarded as selecting it's parent copy randomly among those present in the preceding generation (Nordberg, 2000). Lineages can thus be traced backwards in time until they coalesce in the most recent common ancestor (MRCA) from which the gene was derived, essentially a generalisation of Malécot's concept of *identity by descent* which has been central to genealogical approaches to population genetics (Nagylaki, 1989). A population-level map can therefore be regarded as a map of haplotypes present in the MRCAs from which the population was derived. These haplotypes are in disequilibrium due to limited recombination caused by a small number of generations from the MRCAs. Furthermore, linkage disequilibrium may be maintained in haplotype regions contributing to phenotypic traits due to selection, making these regions more accessible to mapping. The coalescent can be fully exploited if the ancestral genotypes giving rise to the mapped population are marker-typed at the same loci. The origin of chromosome bins or haplotypes in linkage disequilibrium within the population can then be traced, yielding valuable information on population history. Examining the distribution of linkage groups in disequilibrium within the population and within/between pedigree lineages may also be able to shed light on other potential sources of disequilibrium, such as population structure or admixture.

To emphasise once again, the objective of the work described below is to create a population-level map of molecular markers in disequilibrium within a sugarcane breeding population, in order to increase the efficiency of molecular breeding for pest and disease resistance. This is somewhat different to using LD for fine mapping of single genes for map-based cloning, which is the goal of many LD studies in humans, as described by Zhang *et al.* 2002. When using molecular markers in a GAP breeding strategy to pyramid desirable alleles or selected against undesirable alleles, any cause of marker co-segregation is relevant, including disequilibrium resulting from factors such as population structure. The challenge in molecular breeding is to use all available information within an interpretive framework accounting for possible factors such as phenotype, marker-type and genealogy in order to increase the effectiveness of developing superior sugarcane varieties.

5.2. Materials and methods

5.2.1. Mapping population and ancestral population

The same group of 77 genotypes described in Chapter 3.2.1 was used as a linkage disequilibrium mapping population. One individual, 82F2907, had been dropped from the original population of 78 genotypes as it had been found to be misidentified. In addition to the mapping population, an extra seven genotypes representing some of the ancestral clones important in the genealogies of modern commercial sugarcane breeding germplasm were marker-typed on the same AFLP gels used for the marker identification population. These genotypes were included in order to be able to trace the ancestral origin of haplotypes identified in the mapping population. The list of the genotypes in the mapping population and the ancestral population, along with their parents, grandparents and great-grandparents is given in Table 5.1. As genealogical relationships may be important in the interpretation of the LD map within a coalescent framework, a family tree showing the pedigree of the 'NCo' varieties and the position of the seven ancestral genotypes marker-typed is given in Figure 5.1. Unfortunately two important clones within the pedigree, *viz*, CO213 and CO421, are not represented in the ancestral population.

The same molecular marker data set used for QTA identification as described in Chapter 3.2.2 and 3.2.3 was used for detection of disequilibrium and mapping. RFLP data was not available for the seven ancestral clones, however. In summary, the molecular data consisted of 275 RFLP markers scored on the mapping population only, and 1057 AFLP markers scored across both the mapping and ancestral populations. As described in Chapter 3.3.2, analysis of marker frequency and genetic diversity did not detect significant population stratification or structure within the mapping population. The dataset was analysed as a single population, as there was no prior information suggesting the risk of detecting false linkage disequilibrium through population admixture.

5.2.2. Map construction

Association between all pairs of markers in the mapping population was estimated using the r statistic with a threshold value of $r > 0.48$, as described in Chapter 4.3.1. The seven ancestral clones were not included in the analysis. Groups of markers showing significant associations were then ordered and assembled into haplotypes using the 'branch and bound' method as described in Chapter 4.2.5, using a recursive algorithm programmed in GAUSS. This method is guaranteed to find the best order of markers but may require that all possible orders be examined, making it potentially more computationally demanding than other

Table 5.1. Genotypes and their pedigrees for the mapping population and ancestral population. P are parents, with letters 1 and 2 referring to female and male parents respectively, GP are grandparents, (e.g. GP21 is the female parent of P2) and GGP are great grandparents (e.g. GGP121 is the female parent of GP12). PC in the parent column refers to polycross, where the male parent is unknown. ? indicates parent is unknown, while a blank indicates an ancestral *Saccharum* clone in the previous generation.

Variety	P1	P2	GP11	GP12	GP21	GP22	GGP111	GGP112	GGP121	GGP122	GGP211	GGP212	GGP221	GGP222
NCo293	CO421	CO312	POJ2878	CO285	CO213	CO244	POJ2364	EK 28	St.mauritius	S.spont	POJ213	Kansar	POJ213	CO205
NCo339	CO421	CO312	POJ2878	CO285	CO213	CO244	POJ2364	EK 28	St.mauritius	S.spont	POJ213	Kansar	POJ213	CO205
NCo376	CO421	CO312	POJ2878	CO285	CO213	CO244	POJ2364	EK 28	St.mauritius	S.spont	POJ213	Kansar	POJ213	CO205
N52/219	NCO339	NM214	CO421	CO312	M168/32	CO301	POJ2878	CO285	CO213	CO244	Uba marot	POJ2878	CO213	POJ1499
NM214	M168/32	CO301	Uba marot	POJ2878	CO213	POJ1499	S.off	S.spont	POJ2364	EK 28	POJ213	Kansar	POJ385	POJ181
68W1049	NCO310	PC	CO421	CO312	?	?	POJ2878	CO285	CO213	CO244	?	?	?	?
73L1295	N55/805	62L85	NCO310	CO301	CP36/85	NCO382	CO421	CO312	CO213	POJ1499	POJ2725	CP1165	POJ2725	CO301
74M0659	CP57/614	?	CL47/143	CP53/17	?	?	CL41/142	CL41/114	F36/819	CP48/126	?	?	?	?
75E0247	CP59/22	N52/451	CP52/114	CP53/18	NM222	CO285	CP43/64	CP33/372	F36/819	CP48/126	M168/32	CO301	St.mauritius	S.spont
75E1293	NCO376	F152	CO421	CO312	H32-8560	PT43-52	POJ2878	CO285	CO213	CO244	CO213	POJ2878	?	?
75L1157	N55/805	CO312	NCO310	CO301	CO213	CO244	CO421	CO312	CO213	POJ1499	POJ213	Kansar	POJ213	CO205
75L1463	CO331	B42231	CO213	CO214	B3354	CP28/11	POJ213	Kansar	St.mauritius	M4600	?	?	CO281	US1694
76H0333	NCO293	N54/221	CO421	CO312	NCO310	G255	POJ2878	CO285	CO213	CO244	CO421	CO312	POJ2725	CP38/782
76M1101	N65/1425	PC	N57/1917	CO331	?	?	NM222	CO293	CO213	CO214	?	?	?	?
76M1566	SAIPAN 17	CO285	POJ2725	F28	St.mauritius	S.spont	POJ2364	EK28	POJ161	POJ181				
77F0637	F152	N55/805	H32-8560	PT43-52	NCO310	CO301	CO213	POJ2878	?	?	CO421	CO312	CO213	POJ1499
77F0790	F152	N55/805	H32-8560	PT43-52	NCO310	CO301	CO213	POJ2878	?	?	CO421	CO312	CO213	POJ1499
77L1143	N55/805	J59/3	NCO310	CO301	JA54/309	B42231	CO421	CO312	CO213	POJ1499	?	?	B3354	CP28/11
77L1720	CB40/35	N20	POJ2878	CO290	NCO376	CO285	POJ2364	EK 28	CO221	D74	CO421	CO312	St.mauritius	S.spont
77W1241	F152	N52/214	H32-8560	PT43-52	NCO339	NM214	CO213	POJ2878	?	?	CO421	CO312	M168/32	CO301
78F0909	71W1248	PC	LUNA	CB36/14	?	?	CO281	33MQ157	CO213	?	?	?	?	?
79F1043	CO419	64G79	POJ2878	CO290	NCO376	CO285	POJ2364	EK 28	CO221	D74	CO421	CO312	St.mauritius	SPONT.
79F1855	CO1001	70F2446	CO603	CO743	NCO376	N60/1058	CO421XCO440	CO419			CO421	CO312	NM222	NCO293
79H0181	N58/2239	N6	NCO376	CO419	CO421	CP36/85	CO421	CO312	POJ2878	CO290	POJ2878	CO285	POJ2725	CP1165
79L0181	N55/805	CB40/35	NCO310	CO301	POJ2878	CO290	CO421	CO312	CO213	POJ1499	POJ2364	EK 28	CO221	D74
79L1294	65L0373	PC	NCO310	CO331	?	?	CO421	CO312	CO213	CO214	?	?	?	?
79M0955	CP66/1043	N7	CP52/1	CP57/614	NCO378	CO285	CP29/103	CP33/229	CL47/143	CP53/17	CO421	CO312	St.mauritius	S.spont
80E1496	71E1199	N52/214	CB40/35	CO301	NCO339	NM214	POJ2878	CO290	CO213	POJ1499	CO421	CO312	M168/32	CO301
80F2147	N15	PC	N55/805	CB40/35	?	?	NCO310	CO301	POJ2878	CO290	?	?	?	?
80L0432	N8	J59/3	NCO378	CO285	JA54/309	B42231	CO421	CO312	St.mauritius	S.spont	?	?	B3354	CP28/11
80L0627	N52/451	J59/3	NM222	CO285	JA54/309	B42231	M168/32	CO301	St.mauritius	S.spont	?	?	B3354	CP28/11
80M1257	N17	PC	NCO376	CB38/22	?	?	CO421	CO312	CP27/139	?	?	?	?	?
80W1459	NCO293	PC	CO421	CO312	?	?	POJ2878	CO285	CO213	CO244	?	?	?	?
81L1308	N55/805	N16	NCO310	CO301	NCO376	CO331	CO421	CO312	CO213	POJ1499	CO421	CO312	CO213	CO214
81W0133	69W160	N55/805	NCO376	CB38/39	NCO310	CO301	CO421	CO312	POJ2878	?	CO421	CO312	CO213	POJ1499
81W0447	NCO376	B42231	CO421	CO312	B3354	CP28/11	POJ2878	CO285	CO213	CO244	?	?	CO281	US1694
82F0675	76F3084	NM214	66W1371	Kassoer	M168/32	CO301	POJ2878	NCO293	B.cheribon	Glagah	Uba marot	POJ2878	CO213	POJ1499
83F0448	CO419	CP57/614	POJ2878	CO290	CL47/143	CP53/17	POJ2364	EK 28	CO221	D74	CL41/142	CL41/114	F36/819	CP48/126
84F2753	N18	CP57/614	NCO376	PC	CL47/143	CP53/17	CO421	CO312	?	?	CL41/142	CL41/114	F36/819	CP48/126
85F1628	N11	77F790	CB40/35	NCO293	F152	N55/805	POJ2878	CO290	CO421	CO312	H32-8560	PT43-52	NCO310	CO301
85F2805	77F2099	PC	NCO376	PC	?	?	CO421	CO312	?	?	?	?	?	?

Table 5.1. Continued.

Trt	Variety	P1	P2	GP11	GP12	GP21	GP22	GGP111	GGP112	GGP121	GGP122	GGP211	GGP212	GGP221	GGP222
30	85H0605	73H308	NCO293	NCO339	N53/216	CO421	CO312	CO421	CO312	NCO293	CO453	POJ2878	CO285	CO213	CO244
49	85L1041	N12	PC	NCO376	CO331	?	?	CO421	CO312	CO213	CO214	?	?	?	?
19	85L1056	N12	PC	NCO376	CO331	?	?	CO421	CO312	CO213	CO214	?	?	?	?
13	85L1769	M351/57	PC	NCO310	M99/34	?	?	CO421	CO312	M109/26	Uba marot	?	?	?	?
28	85W1610	76M0547	PC	NCO291	NCO293	?	?	CO421	CO312	CO421	CO312	?	?	?	?
75	87L0329	N18	N14	NCO376	PC	N7	PC	CO421	CO312	?	?	NCO378	CO285	?	?
68	87L1484	N55/805	77L1307	NCO310	CO301	CB40/35	67L855	CO421	CO312	CO213	POJ1499	POJ2878	CO290	NCO376	NCO382
77	88W1323	77F0637	N16	F152	N55/805	NCO376	CO331	H32-8560	PT43-52	NCO310	CO301	CO421	CO312	CO213	CO214
45	N8	NCO378	CO285	CO421	CO312	St.mauritius	S.spont	POJ2878	CO285	CO213	CO244	?	?	?	?
66	N11	CB40/35	NCO293	POJ2878	CO290	CO421	CO312	POJ2364	EK 28	CO221	D74	POJ2878	CO285	CO213	CO244
39	N13	N52/214	NCO293	NCO339	NM214	CO421	CO312	CO421	CO312	M168/32	CO301	POJ2878	CO285	CO213	CO244
59	N14	N7	PC	NCO378	CO285	?	?	CO421	CO312	St.mauritius	S.spont	?	?	?	?
41	N16	NCO376	CO331	CO421	CO312	CO213	CO214	POJ2878	CO285	CO213	CO244	POJ213	Kansar	St.mauritius	M4600
10	N17	NCO376	CB38/22	CO421	CO312	CP27/139	?	POJ2878	CO285	CO213	CO244	D74	US1694	?	?
57	N18	NCO376	PC	CO421	CO312	?	?	POJ2878	CO285	CO213	CO244	?	?	?	?
54	N19	NCO376	CB40/35	CO421	CO312	POJ2878	CO290	POJ2878	CO285	CO213	CO244	POJ2364	EK 28	CO221	D74
26	N20	NCO376	CO285	CO421	CO312	St.mauritius	S.spont	POJ2878	CO285	CO213	CO244				
1	N21	CB38/22	N52/214	CP27/139	?	NCO339	NM214	D74	US1694	?	?	CO421	CO312	M168/32	CO301
35	N22	70E469	N52/219	CO421	N55/805	NCO339	NM214	POJ2878	CO285	NCO310	CO301	CO421	CO312	M168/32	CO301
14	N23	NCO376	N52/219	CO421	CO312	NCO339	NM214	POJ2878	CO285	CO213	CO244	CO421	CO312	M168/32	CO301
34	N24	Q96	75F1463	Q63	Q68	CO475	CO440	TROJAN	CP29/116	POJ2878	CO290	CO421XCO440	CO419	CO360GC	CO361
52	N25	CO62175	N14	CO951	CO416	N7	PC	CO683	CO419	Vellai	CO243	NCO378	CO285	?	?
76	N26	75F1463	69F0607	CO475	CO440	CB40/69	PC	CO421XCO440	CO419	CO360GC	CO361	POJ2878	CO290	?	?
62	N27	NIN2	N52/219	NCO310	CO331	NCO339	NM214	CO421	CO312	CO213	CO214	CO421	CO312	M168/32	CO301
74	N30	77F0637	78F1025	F152	N55/805	CP66/1043	CB40/35	H32-8560	PT43-52	NCO310	CO301	CP52/1	CP57/614	POJ2878	CO290
73	N31	69E0991	PC	C0421	PC	?	?	POJ2878	CO285	?	?	?	?	?	?
71	N32	N12	PC	NCO376	CO331	?	?	CO421	CO312	CO213	CO214	?	?	?	?
69	N33	75L1463	N8	CO331	B42231	NCO378	CO285	CO213	CO214	B3354	CP28/11	CO421	CO312	St.mauritius	S.spont
72	N34	78F1025	68W1049	CP66/1043	CB40/35	NCO310	PC	CP52/1	CP57/614	POJ2878	CO290	CO421	CO312	?	?
3	B42231	B3354	CP28/11	?	?	CO281	US1694	?	?	?	?	POJ213	CO206	POJ213	?
23	CB38/22	CP27/139	?	D74	US1694			Crystalina	?	POJ213	?				
20	CB40/35	POJ2878	CO290	POJ2364	EK 28	CO221	D74	POJ100	Kassoer			POJ213	M2	S.off	NV
2	CO281	POJ213	CO206	B.cheribon	Chunnee	Ashy Mauritius	S.spont								
48	CO285	St.mauritius	S.spont												
6	CP57/614	CL47/143	CP53/17	CL41/142	CL41/114	F36/819	CP48/126	Bourne35/9	CP27/108			F31/962	CP30/24	CP36/105	CP38/34
21	J59/3	JA54/309	B42231	?	?	B3354	CP28/11	?	?	?	?	?	?	CO281	US1694

Table 5.1. Continued.

Trt	Variety	P1	P2	GP11	GP12	GP21	GP22	GGP111	GGP112	GGP121	GGP122	GGP211	GGP212	GGP221	GGP222
	Ancestral clones														
79	CO205	Vellai	S.spont												
80	CO213	POJ213	Kansar	B.cheribon	Chunnee										
81	CO244	POJ213	CO205	B.cheribon	Chunnee	Vellai	S.spont								
82	CO312	CO213	CO244	POJ213	Kansar	POJ213	CO205	B.cheribon	Chunnee			B.cheribon	Chunnee	Vellai	Chunnee
86	POJ2364	POJ100	Kassoer	B.borneo	Loethers	B.cheribon	Glagah								
87	POJ2878	POJ2364	EK 28	POJ100	Kassoer	EK2	POJ100	B.borneo	Loethers	B.cheribon	Glagah	Lahaina	B.fiji	B.borneo	Loethers
84	Coimbatore														
		S.officinarum					**S.barberi / natural hybrid**			**S.spontaneum**					
		B.cheribon	Lahaina	St.mauritius	Ashy mauritius		Kassoer	Kansar		Glagah					
		B.borneo	Loethers	Crystalina	POJ100		Chunnee			Coimbatore					
		B.fiji	Vellai	EK2	D74		Uba marot			S.spont - unknown clone					

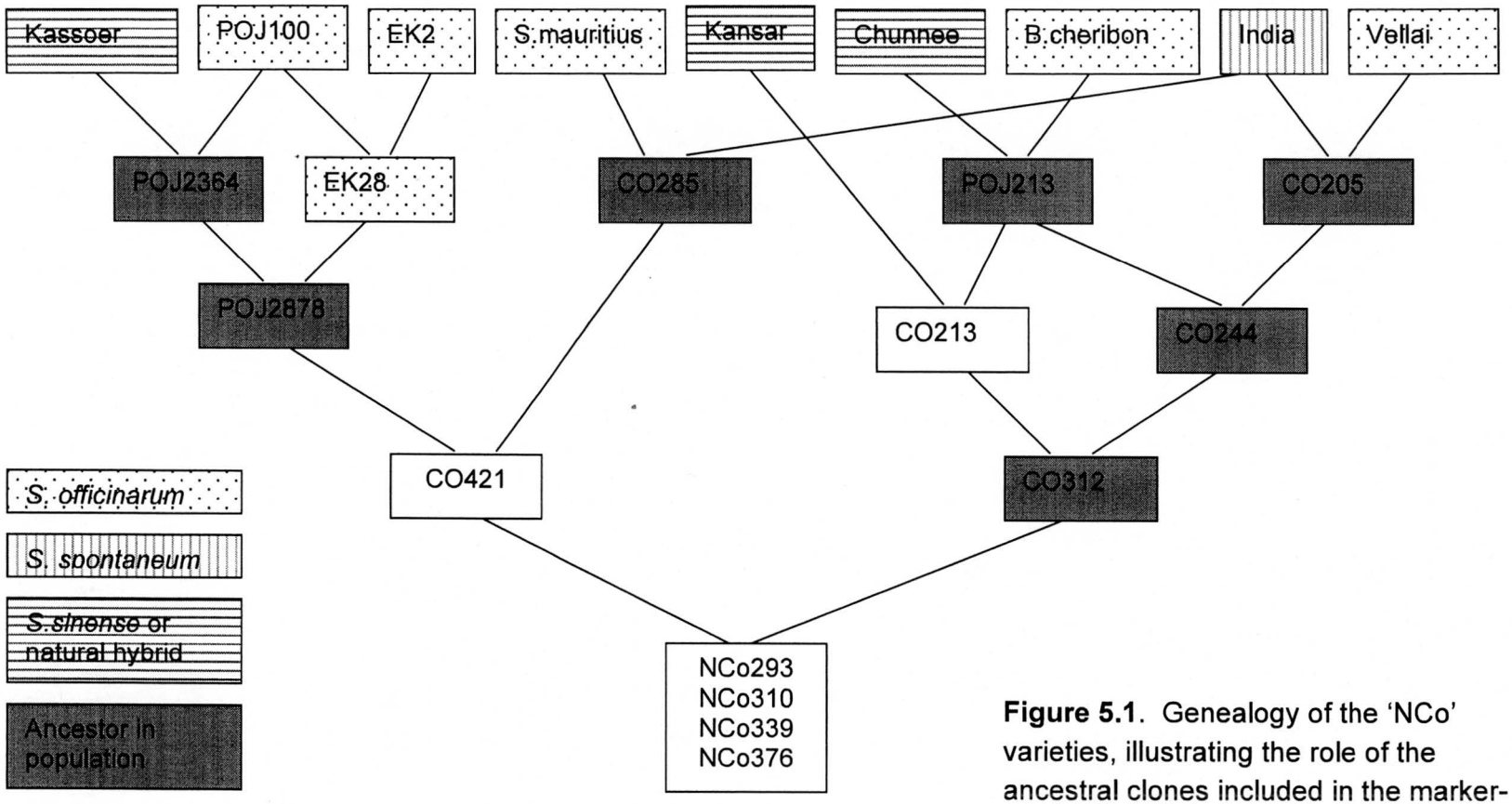


Figure 5.1. Genealogy of the 'NCo' varieties, illustrating the role of the ancestral clones included in the marker-typed population.

methods such as seriation or simulated annealing (Weir, 1996). The GAUSS code for estimating allelic association and assembling markers into haplotypes is given in Appendix A.

Marker data for each genotype within the mapping population as well as the ancestral population was superimposed on the linkage disequilibrium population map to give a map of each genotype. Marker-trait associations identified for response to smut and eldana as described in Chapter 3.3.3 were included on the map, to reveal haplotypes containing putative QTAs for resistance or susceptibility.

5.2.3. Identification of linkage groups for use in breeding

Markers included on the map were used to repeat the stepwise multiple regression to select subsets of six markers ascribing the maximum variation in resistance to smut and eldana respectively, as described in Chapter 3.3.4. It was necessary to repeat the multiple regression as some of the markers shown in Tables 3.10b and 3.11b (Chapter 3.3.4) were not mapped on the LD map. When selecting the markers for smut resistance, the dataset was restricted to markers showing a statistically significant association with smut, but where the correlation with eldana was less than |0.15|. A similar restriction was applied on markers when selecting the subset for eldana resistance. This was to prevent selecting markers associated with resistance to one trait but having an undesirable correlated selection response with the second trait. For each trait, the linkage groups associated with the set of markers were then extracted from the map. The presence of individual linkage groups in the mapping population and in the ancestral population was examined in order to infer the likely ancestral source of linkage groups of interest.

5.3. Results

5.3.1. Mapping and marker-trait associations

A total of 1693 marker pairs involving 841 different markers were identified as being significantly associated, or in linkage disequilibrium. The 1693 marker-pairs were assembled into 231 linkage groups (LGs) or haplotypes using the 'branch and bound' algorithm. The distribution of numbers of markers to linkage groups is shown in Table 5.2. The majority of linkage groups had only two markers in linkage disequilibrium, but some regions containing multiple markers were also identified. Because the *r* statistic has no consistent interpretation in terms of distance in centi-Morgans, no attempt was made to estimate the extent of linkage disequilibrium contained within the haplotypes.

Figure 5.2. Population-level map of markers displaying linkage disequilibrium within genotypes in the SASRI breeding population. A black square represents the presence of the marker, and a white square indicates absence. A grey square represents missing data. The frequency of each marker is shown, along with the correlation coefficient with smut and eldana rating. A negative correlation indicates resistance, while a positive correlation indicates suseptibility. Markers significant at r > |0.25| are highlighted in bold.

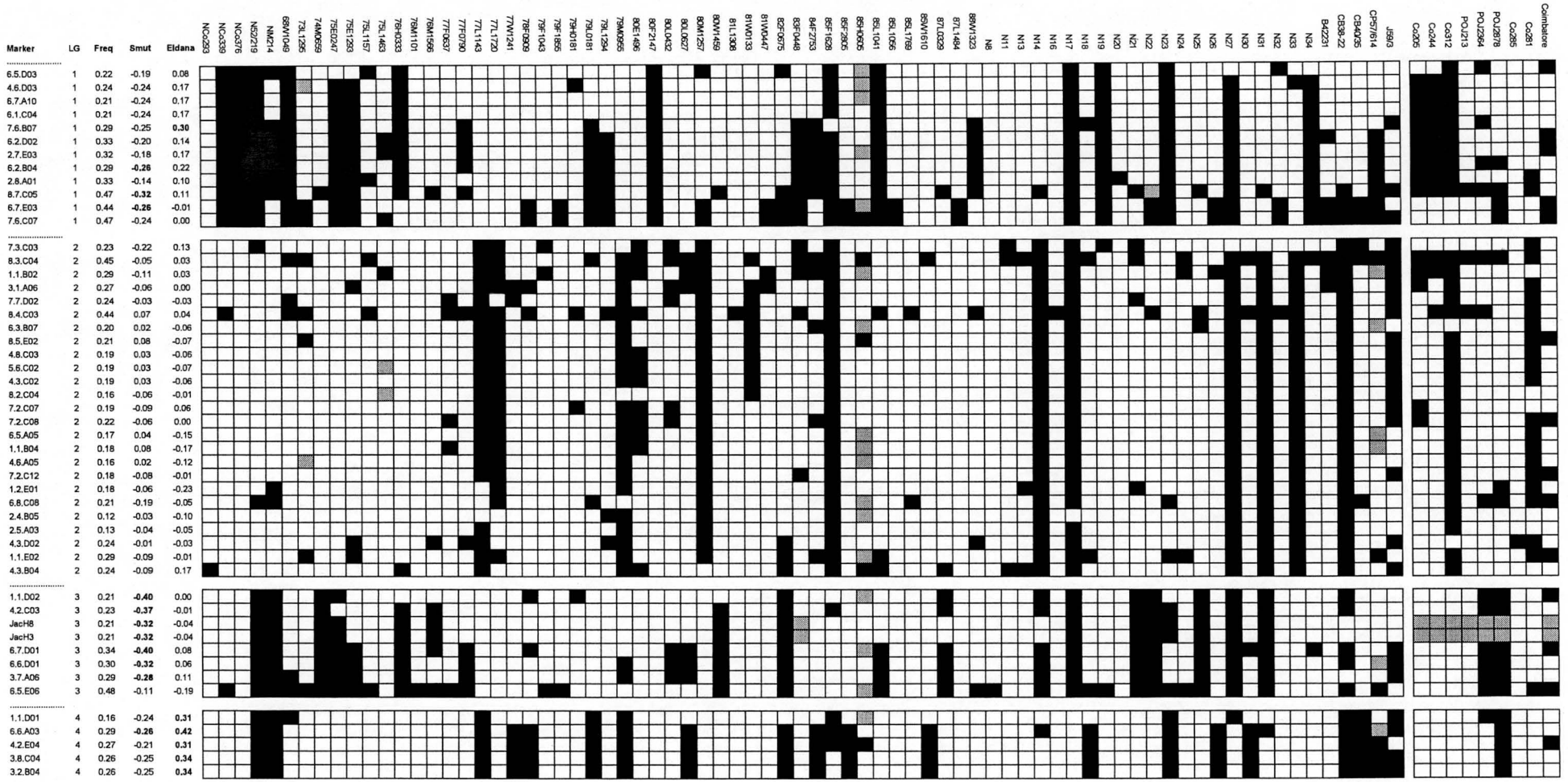


Figure 5.2. Continued.

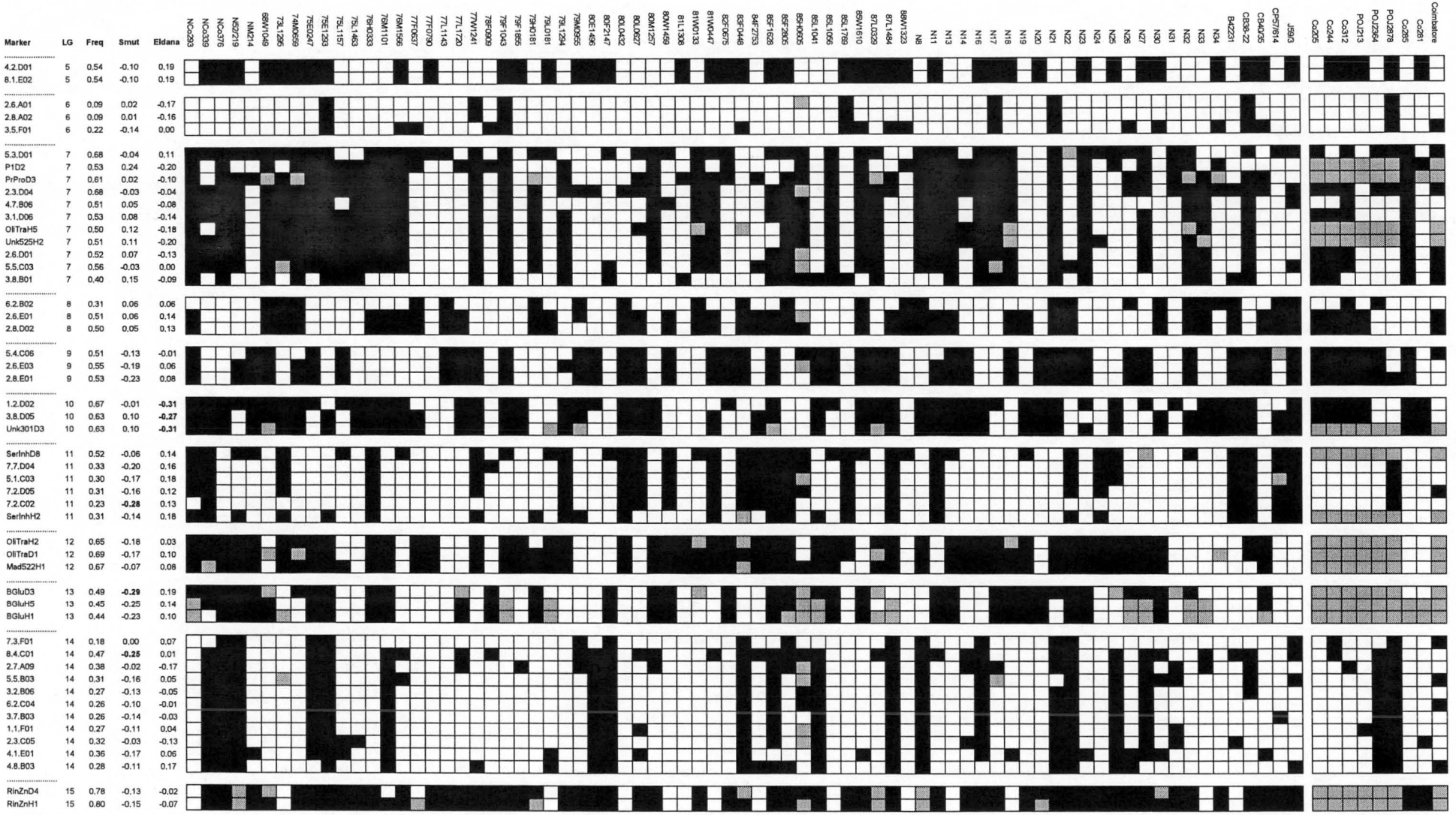


Figure 5.2. Continued.

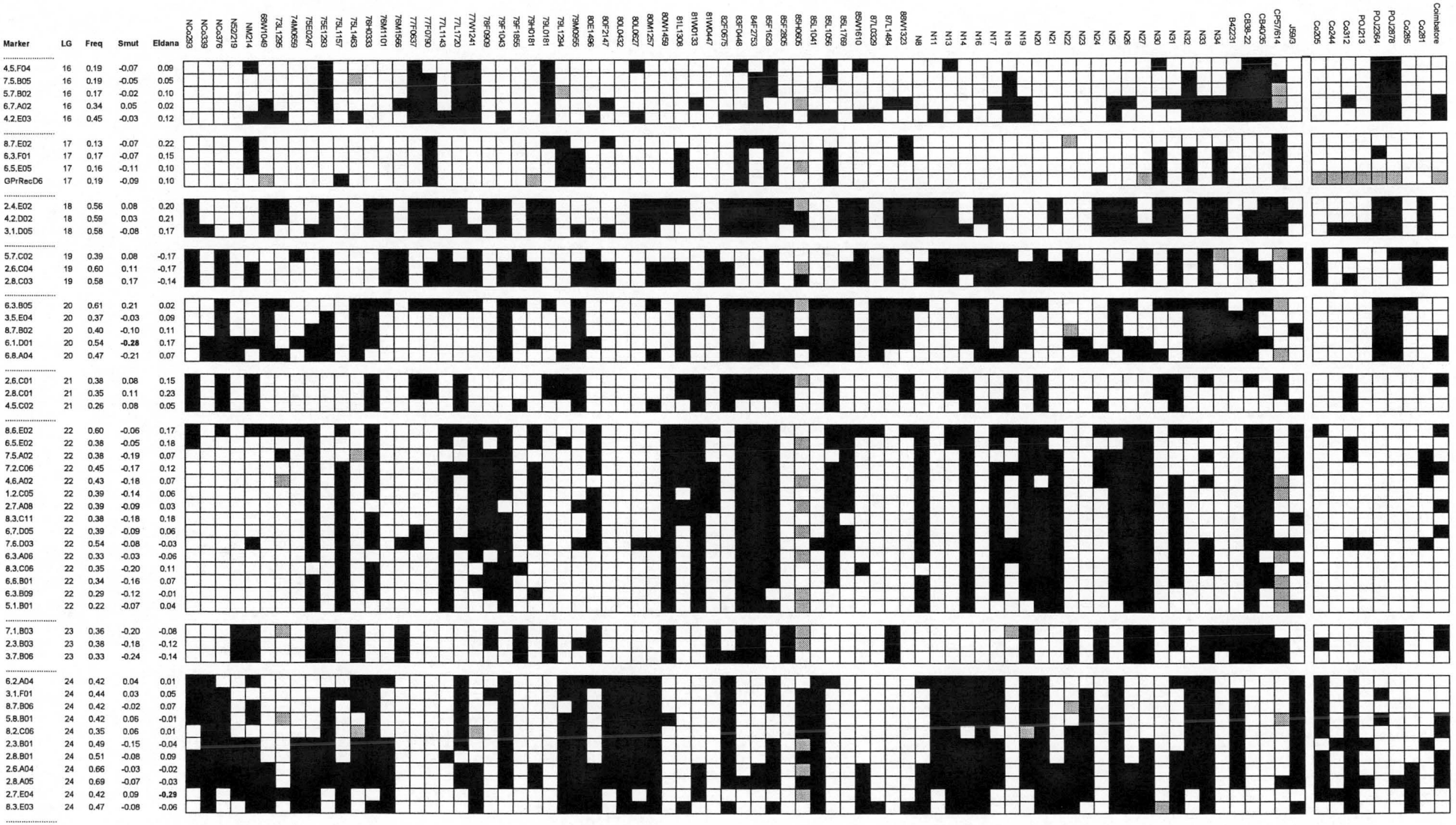


Figure 5.2. Continued.

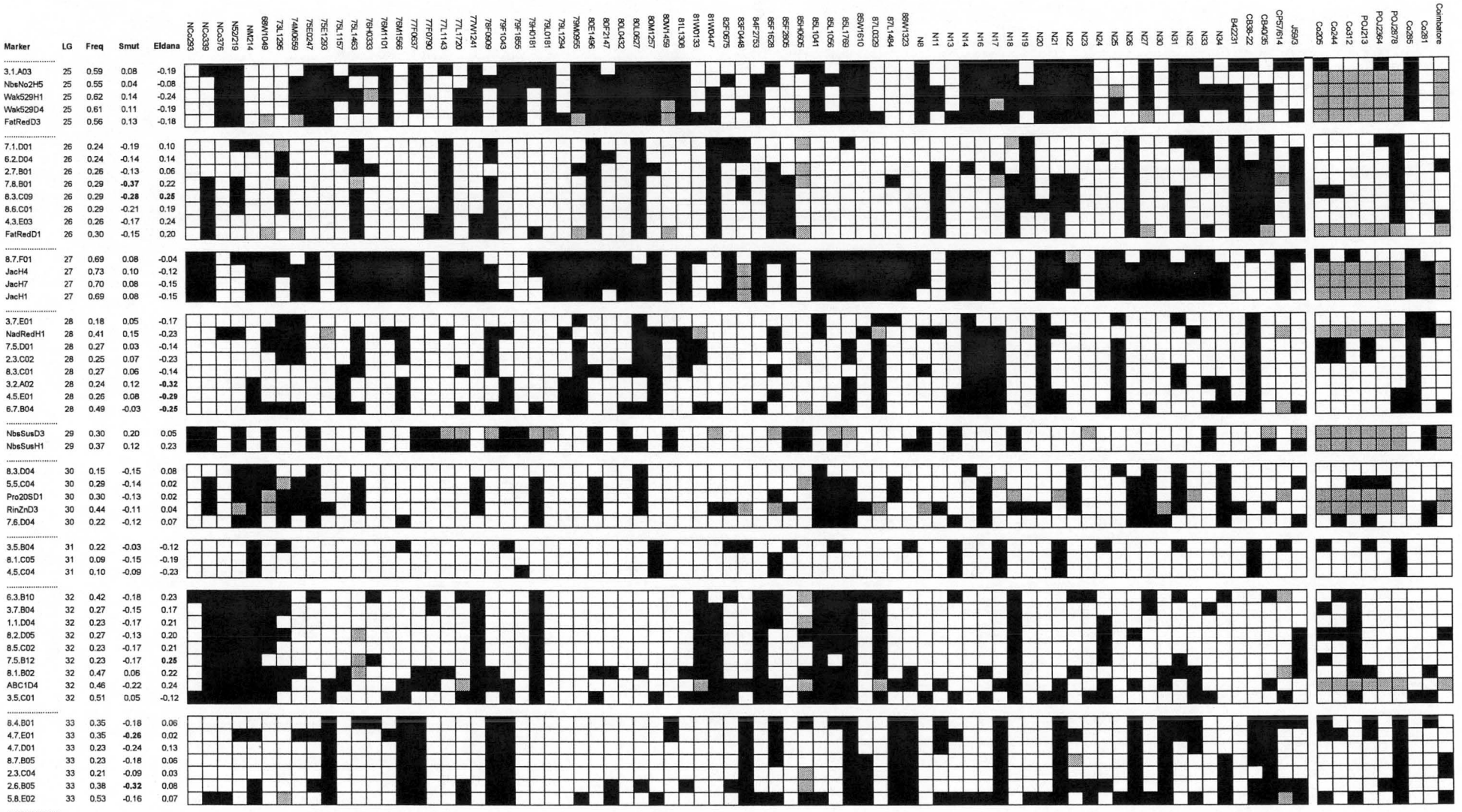


Figure 5.2. Continued.

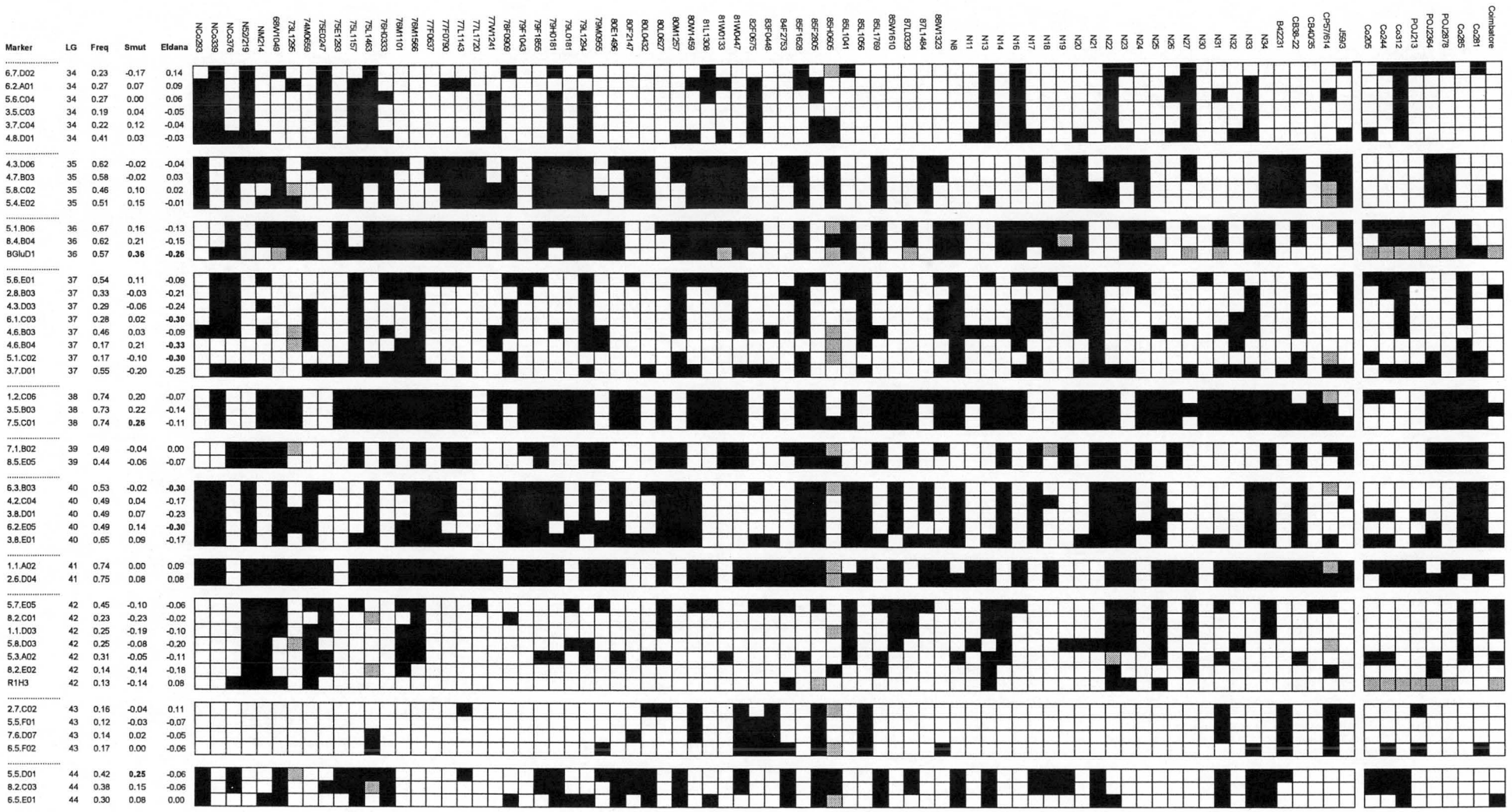


Figure 5.2. Continued.

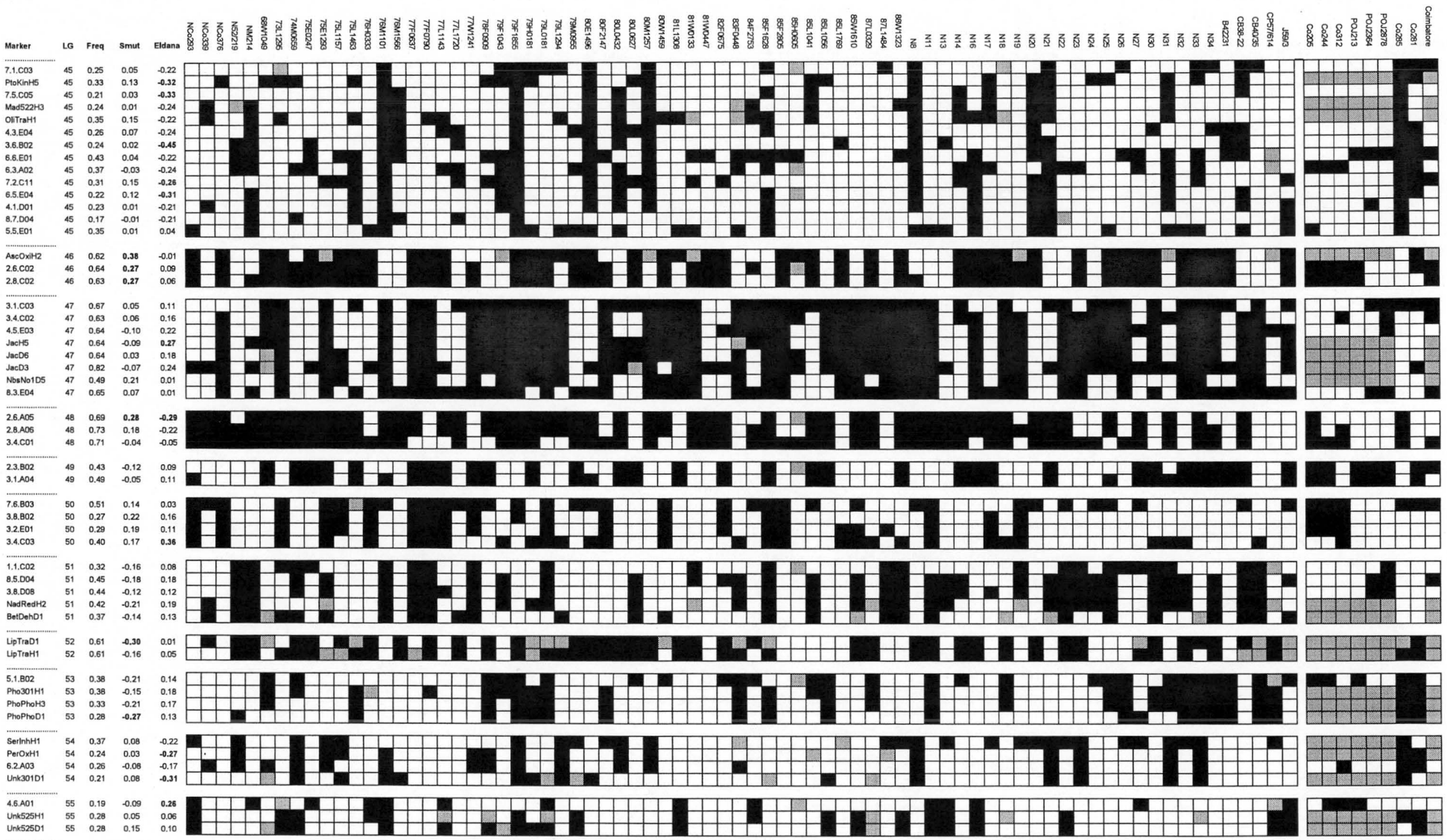


Figure 5.2. Continued.

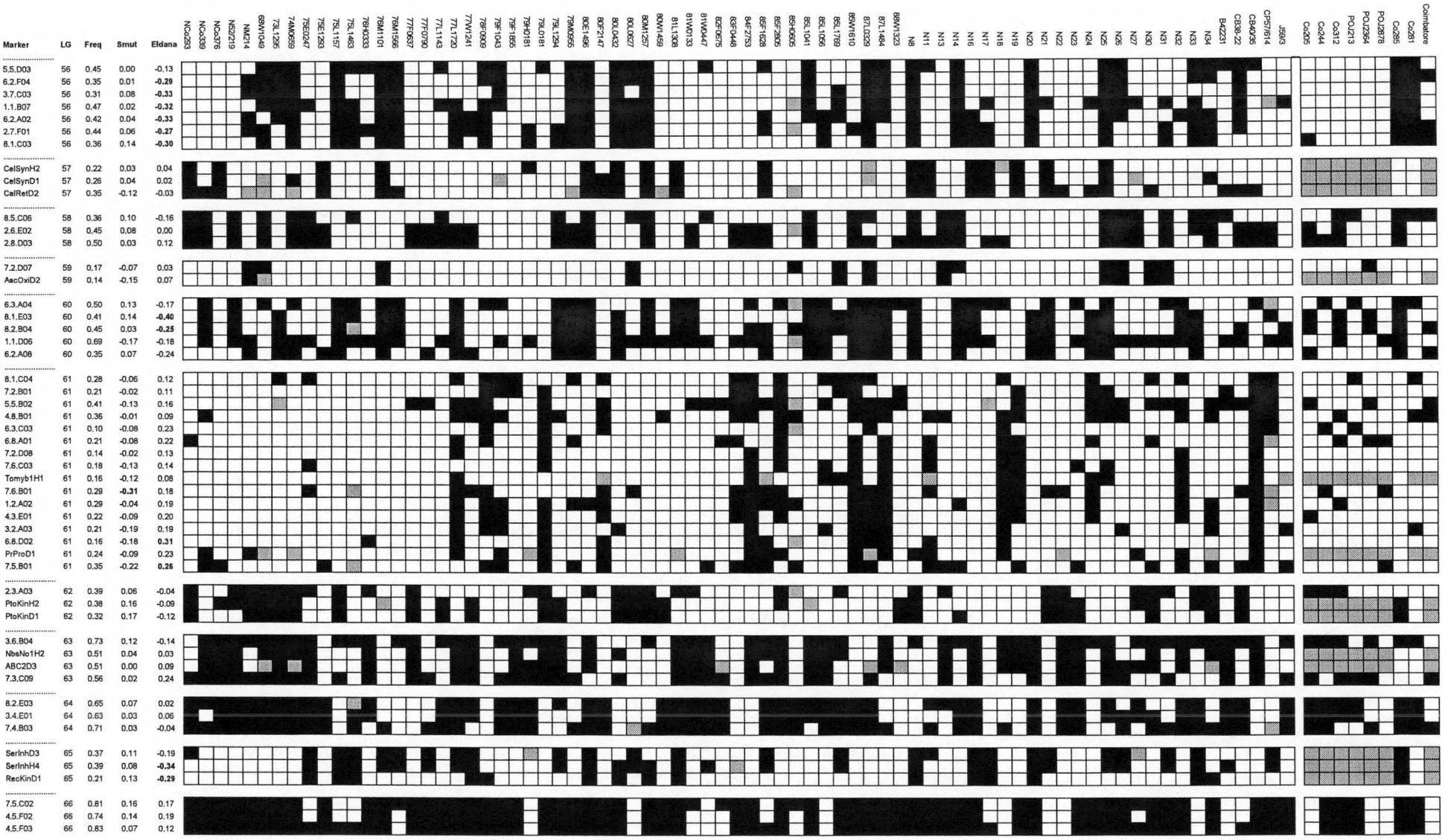


Figure 5.2. Continued.

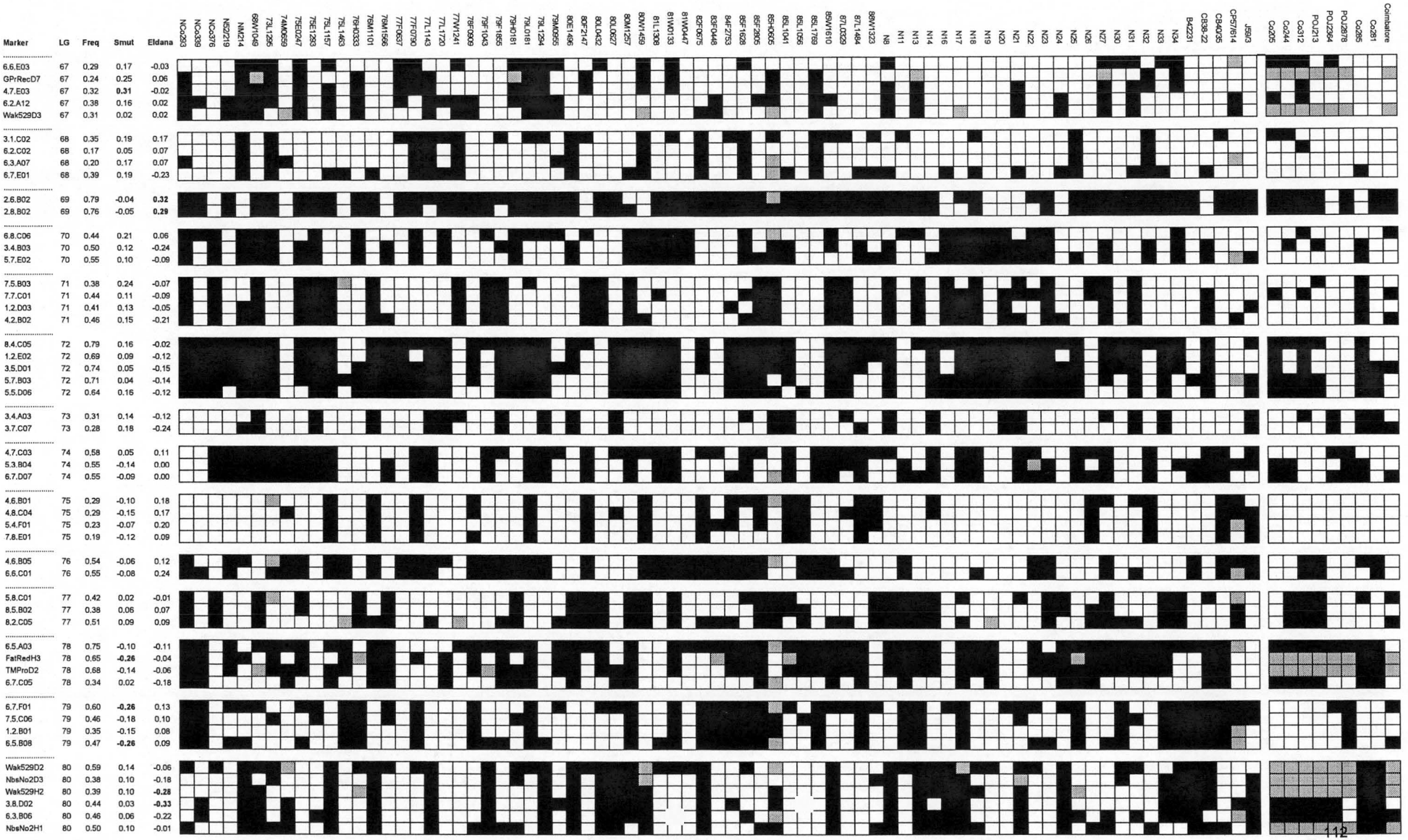


Figure 5.2. Continued.

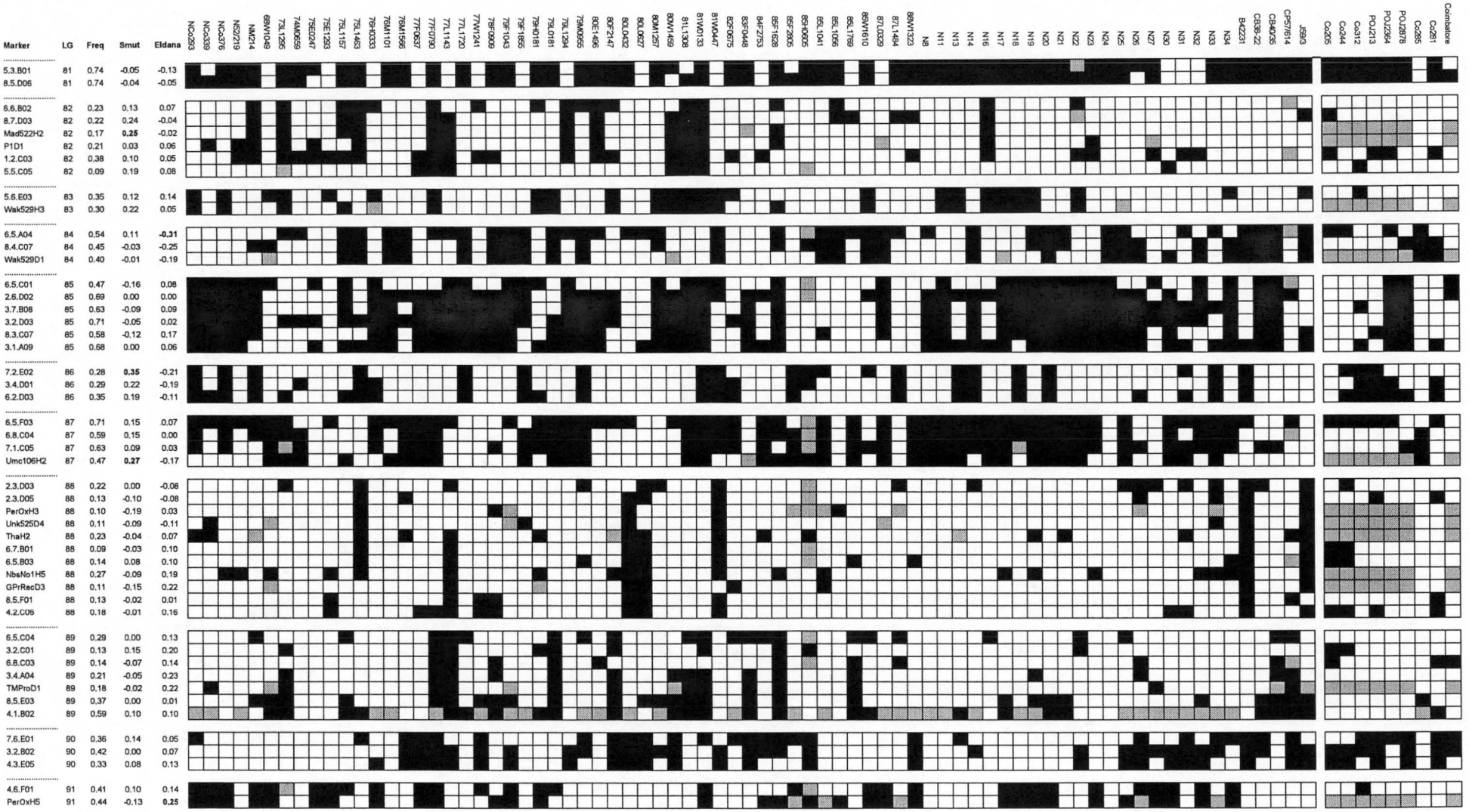


Figure 5.2. Continued.

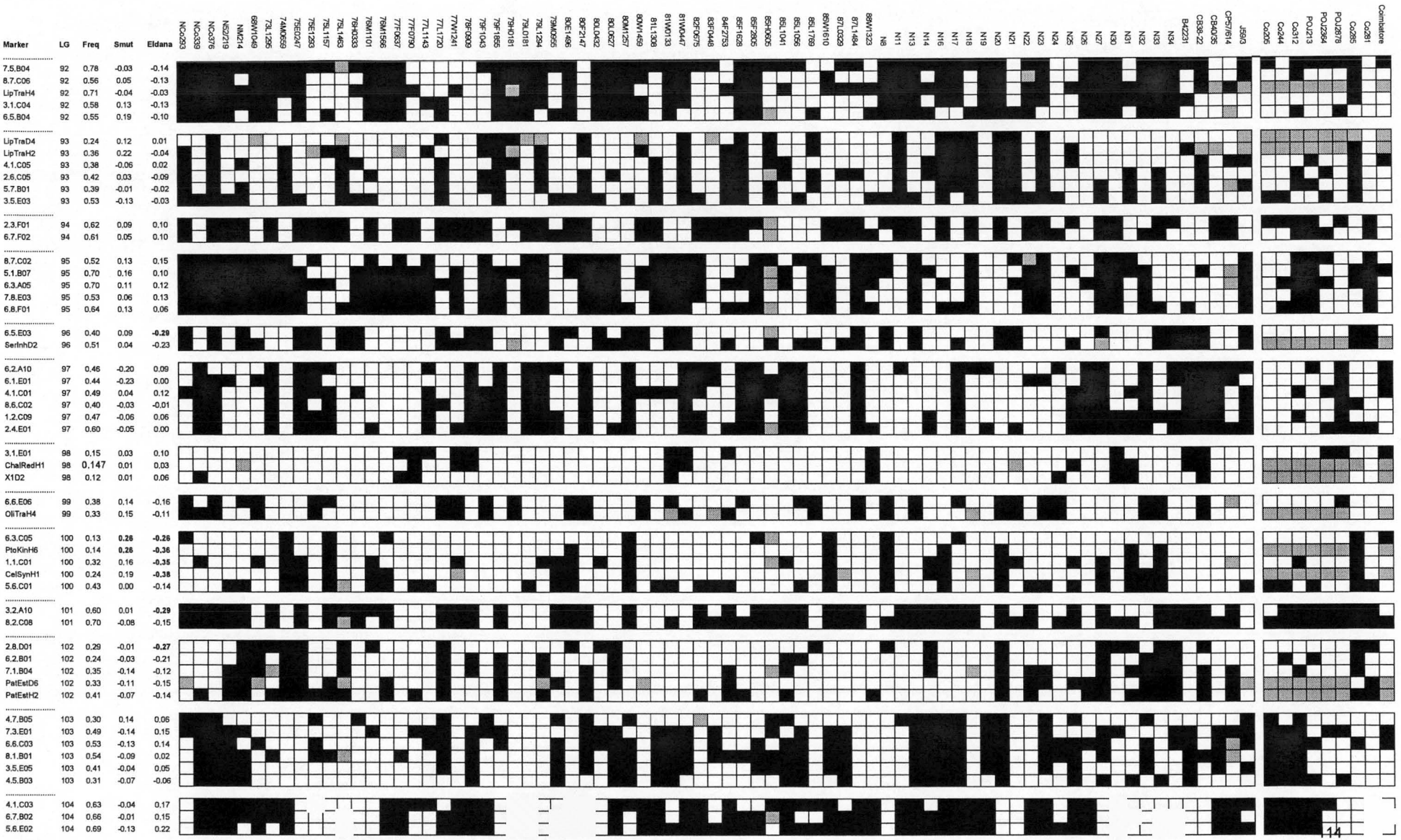


Figure 5.2. Continued.

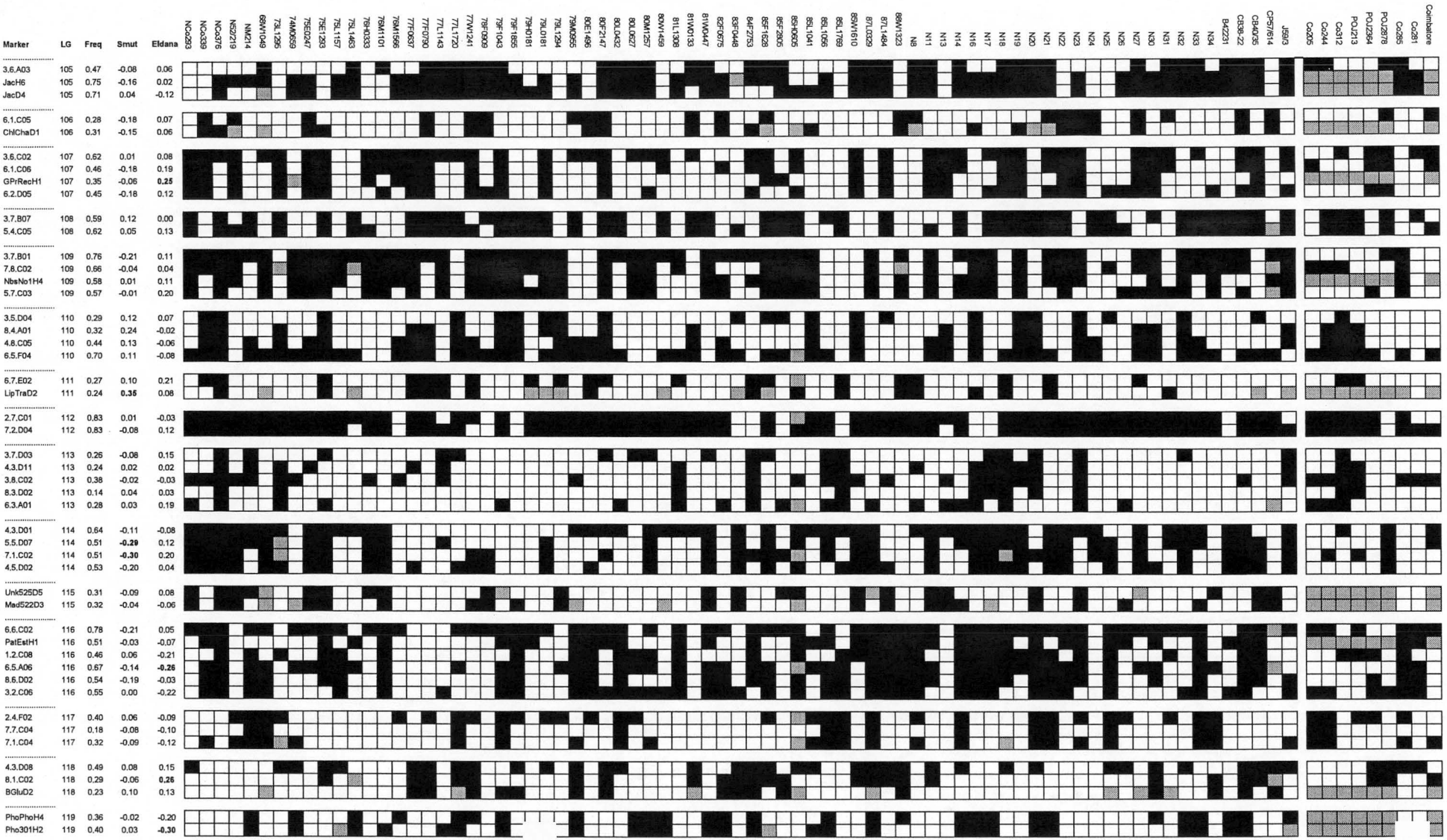


Figure 5.2. Continued.

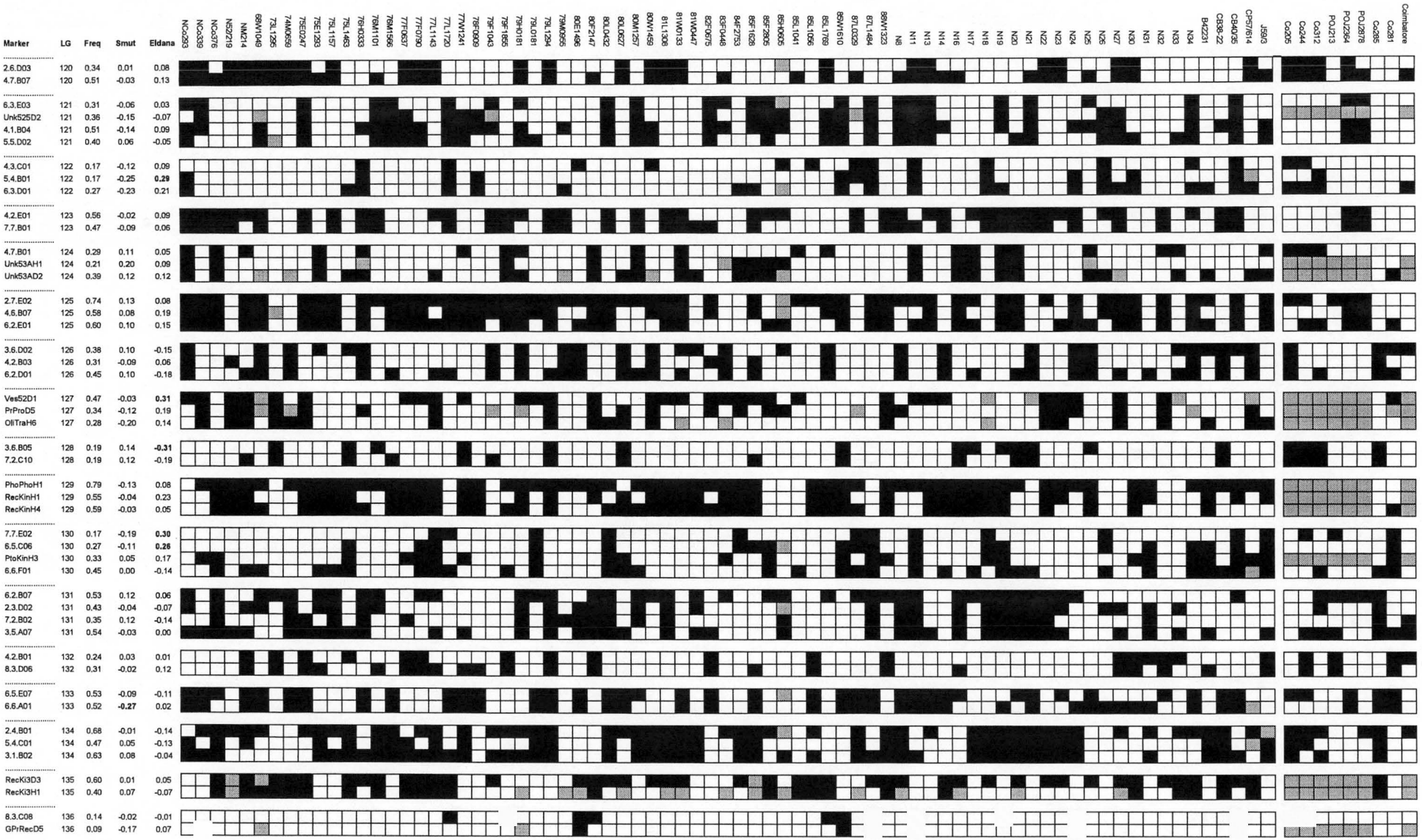


Figure 5.2. Continued.

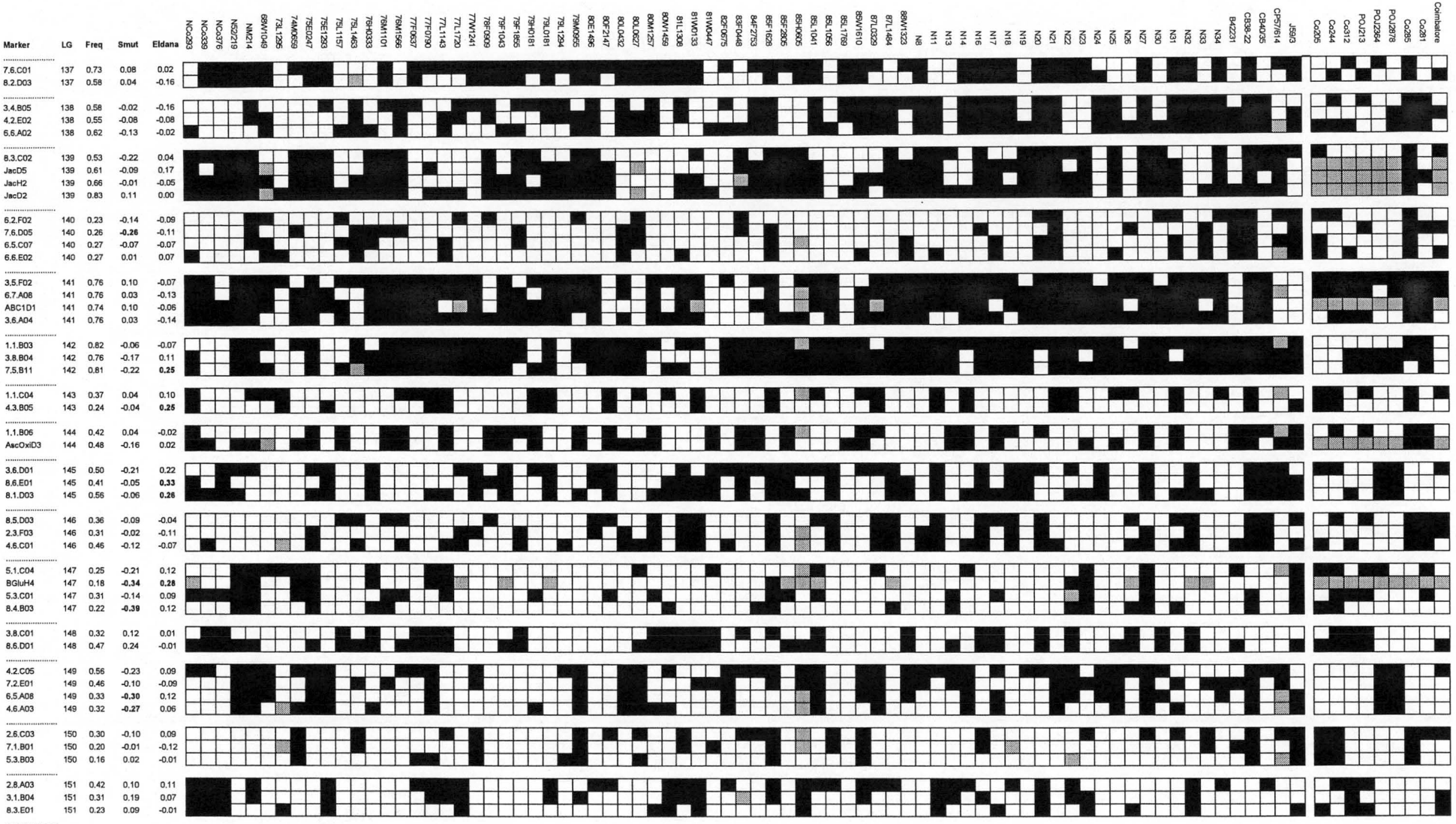


Figure 5.2. Continued.

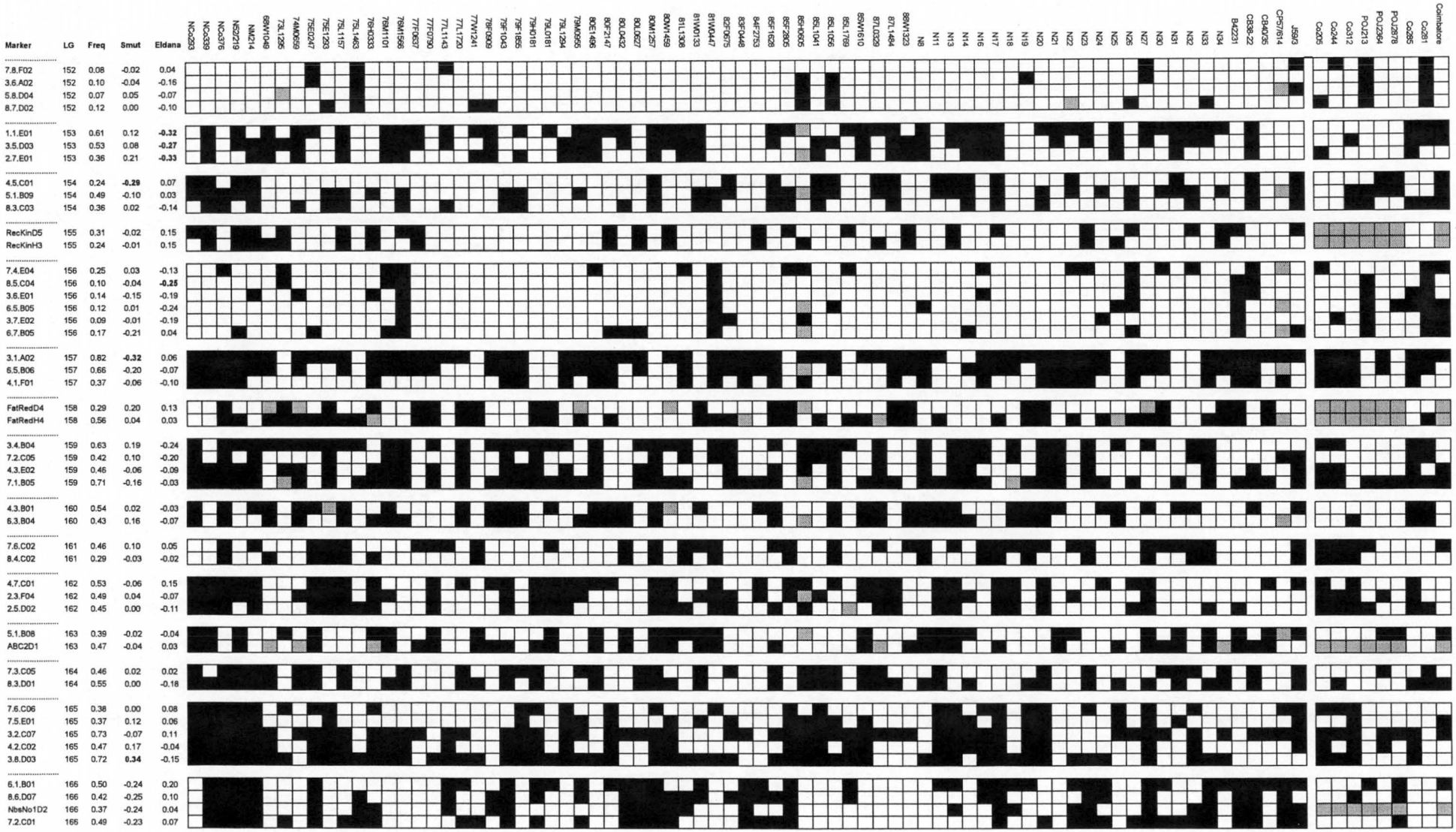


Figure 5.2. Continued.

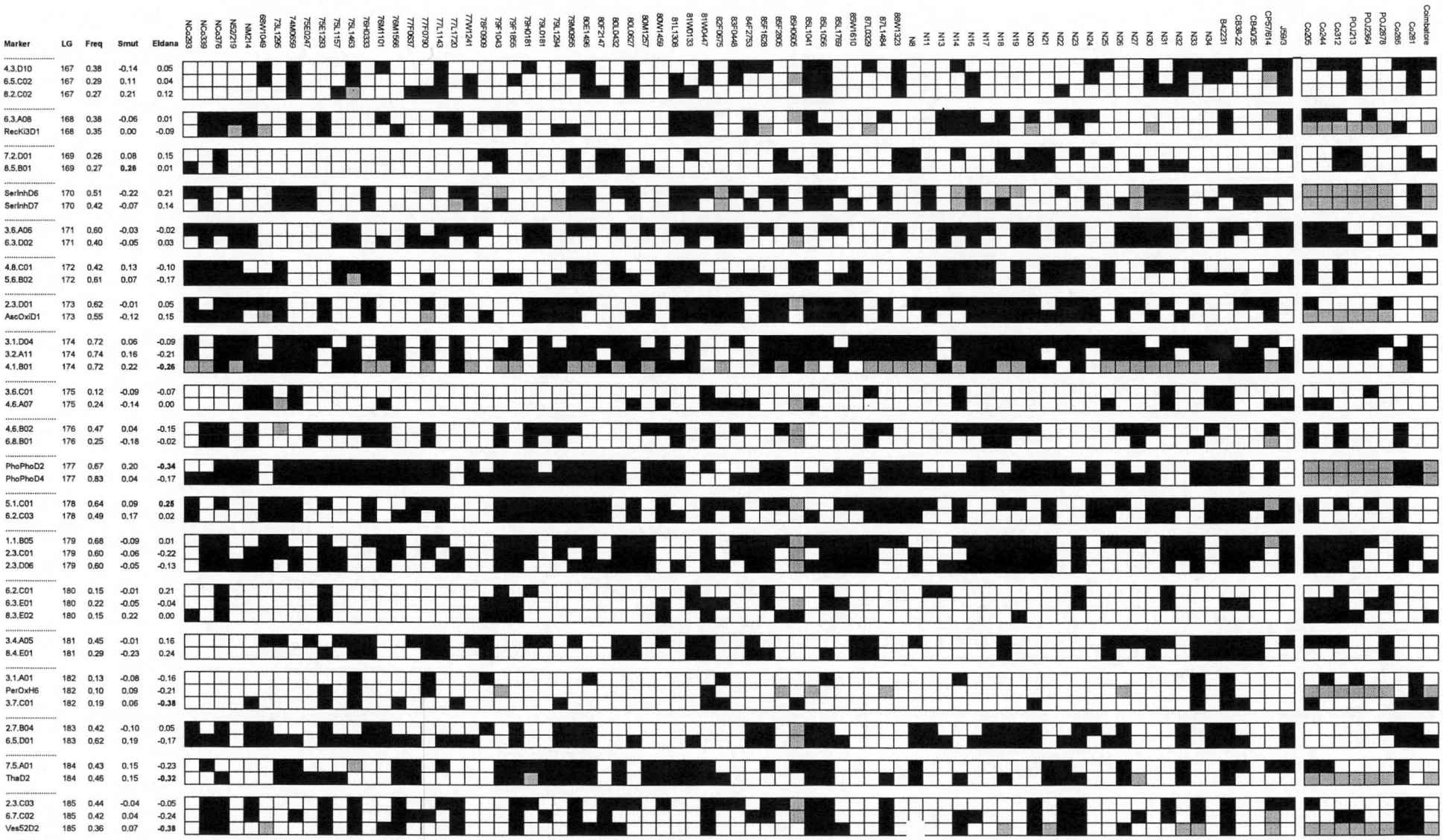


Figure 5.2. Continued.

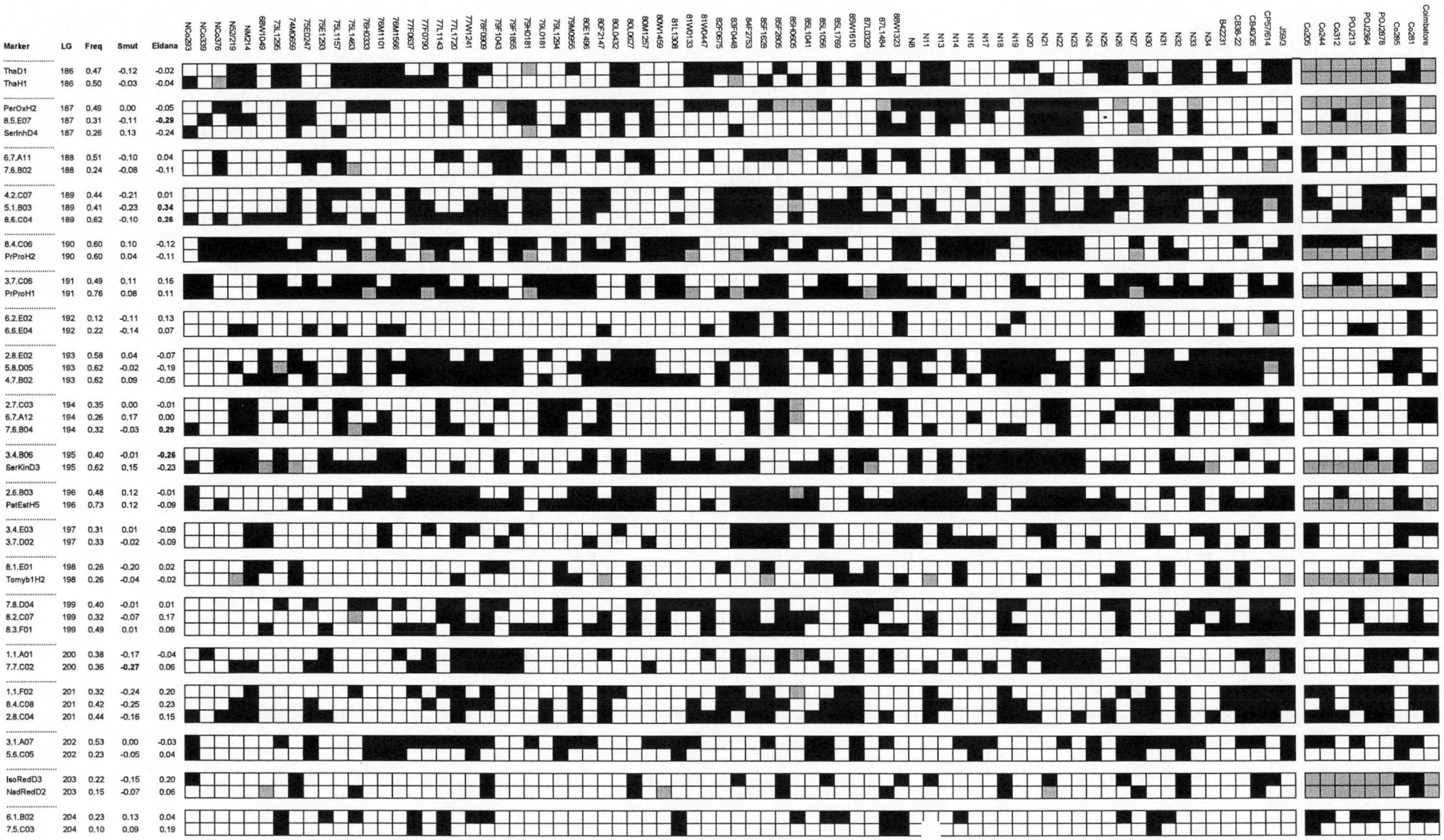


Figure 5.2. Continued.

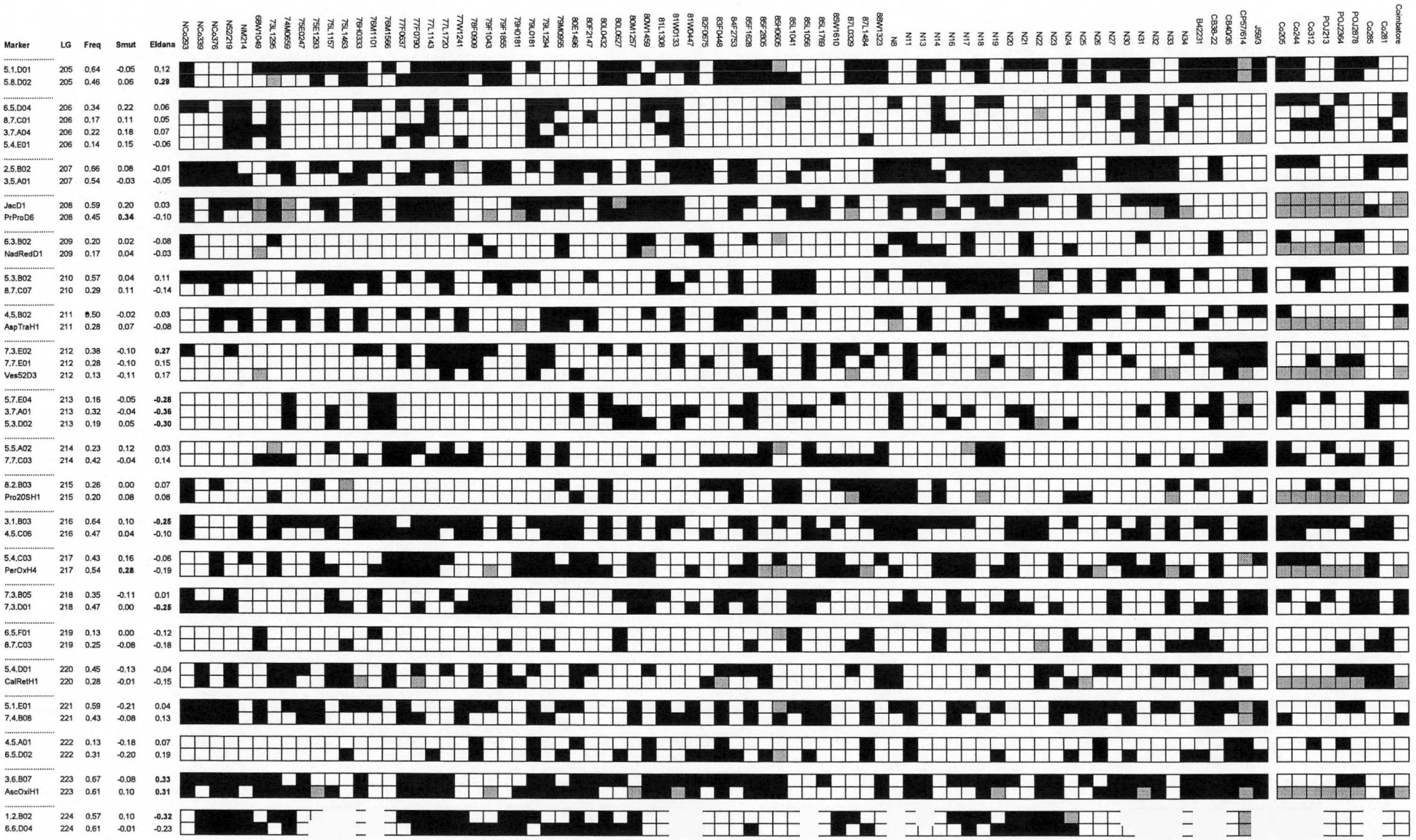


Figure 5.2. Continued.

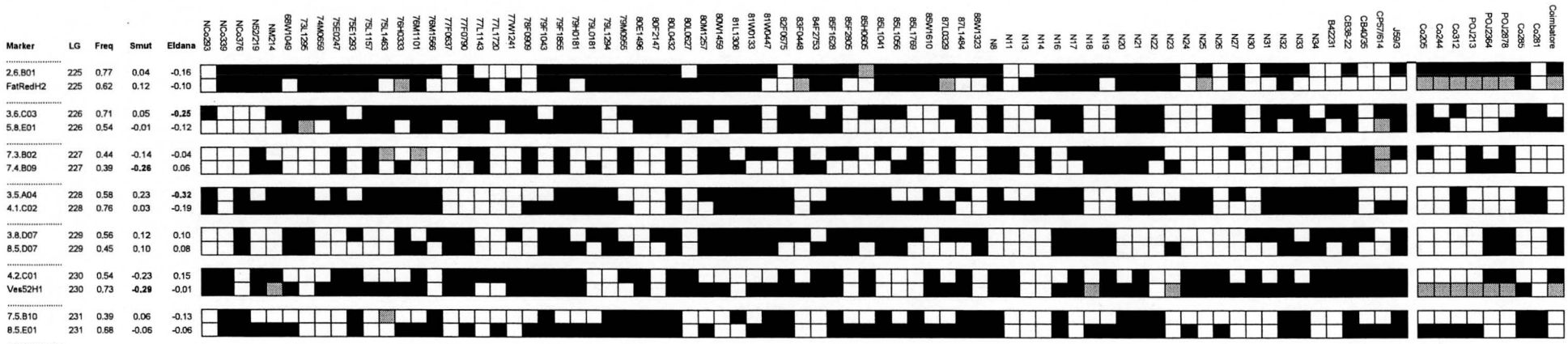


Table 5.2. Distribution of LGs with respect to the number of markers

Number of markers in LG	Number of linkage groups
2	91
3	65
4	30
5	15
6	11
7	4
8	5
9	1
10	0
11	4
12	1
13	0
14	1
15	1
16	1
25	1

The results from Chapter 4.3.5 (Table 4.7) suggest that many LGs are likely to be in the range of 10 cM, while the larger linkage groups may be of the order of 50 cM. The resulting map of 231 haplotypes is shown in Figure 5.2. Although the genotypes CO281 and CO285 were part of the mapping population, they are shown in Figure 5.2 in the group of ancestral clones, as they are both early-generation hybrids occurring within the genealogy of modern cultivars.

An attempt was made to relate the LD map to the map of variety R570 by comparing the markers scored against the gel images from the R570 mapping population through the TropGENE database (http://tropgenedb.cirad.fr). This proved to be very difficult due to the quality of the scanned images and the variation in migration distances between different gel sets, so haplotypes from the LD map given in Figure 5.2 could not be assigned to Homology Groups based on the R570 map.

Significant associations identified with smut or eldana resistance rating as described in Chapter 3.3.3 are included in Figure 5.2. Of the 64 smut-associated markers described previously, 57 were mapped in 40 linkage groups, and 7 significant markers remained unmapped. For eldana, 99 of the 115 identified markers were mapped into 62 linkage

groups, and 16 markers remained unmapped. The number of mapped markers associated with resistance or susceptibility for smut and eldana are shown in Table 5.3. The number of linkage groups the markers are located in are shown in parentheses. Part of the benefit of using a map-based approach to molecular breeding as opposed to a simple marker-trait association approach is evident in Table 5.3, where 63 individual associations for eldana resistance have been 'collapsed' into 38 linkage groups, making the choice of LGs for use in breeding simpler.

Table 5.3. Number of mapped markers (and LGs) associated with smut and eldana at r > |0.25|

	Smut	**Eldana**
Resistant	39 (25)	63 (38)
Susceptible	18 (15)	36 (24)

In order to investigate the effect of individual markers on the correlated response to smut and eldana, the marker-trait association threshold was dropped to r > |0.20|, and mapped markers exceeding this value for smut and eldana were extracted. These are shown in Table 5.4. Thirty one markers in 23 linkage groups were associated with both smut and eldana at r > |0.20|, and in every case the sign of the association was opposite – i.e. markers associated with resistance to one trait were associated with susceptibility to the other. Two markers associated with both smut and eldana (BGluD1 and BGluH4) were derived from an RFLP probe with homology to genes encoding beta-glucosidase. This class of genes has been implicated cyanogenesis, the release of toxic hydrogen cyanide (HCN) from damaged tissue, which is recognized as a general plant defense mechanism against herbivory (Nahrstedt, 1985). For example, Ballhorn *et al.*, (2006) measured increase enzyme activity of beta-glucosidase and subsequent release of HCN in from bean leaves in response to damage from spider mites. If this gene were present in one homologous set of chromosomes, different markers (or linkage groups) would represent different alleles at the locus. From the map, all LGs with BGlu markers were extracted, and shown in Table 5.5.

Table 5.4. Markers associated with both smut and eldana at r > |0.20|. Markers with r > |0.28| are highlighted in bold.

Marker	Linkage group	Marker frequency	Smut	Eldana
7.6.B07	1	0.29	-0.25	**0.30**
6.2.B04	1	0.29	-0.26	0.22
1.1.D01	4	0.16	-0.24	**0.31**
6.6.A03	4	0.29	-0.26	**0.42**
4.2.E04	4	0.27	-0.21	**0.31**
3.8.C04	4	0.26	-0.25	**0.34**
3.2.B04	4	0.26	-0.25	**0.34**
7.8.B01	26	0.29	**-0.37**	0.22
8.3.C09	26	0.29	-0.28	0.25
ABC1D4	32	0.46	-0.22	0.24
BGluD1	36	0.57	**0.36**	-0.26
4.6.B04	37	0.17	0.21	**-0.33**
2.6.A05	48	0.69	0.28	**-0.29**
7.5.B01	61	0.35	-0.22	0.26
7.2.E02	86	0.28	**0.35**	-0.21
6.3.C05	100	0.13	0.26	-0.26
PtoKinH6	100	0.14	0.26	**-0.36**
5.4.B01	122	0.17	-0.25	**0.29**
6.3.D01	122	0.27	-0.23	0.21
7.5.B11	142	0.81	-0.22	0.25
3.6.D01	145	0.50	-0.21	0.22
BGluH4	147	0.18	**-0.34**	**0.28**
2.7.E01	153	0.36	0.21	**-0.33**
6.1.B01	166	0.50	-0.24	0.20
SerInhD6	170	0.51	-0.22	0.21
4.1.B01	174	0.72	0.22	-0.26
8.4.E01	181	0.29	-0.23	0.24
5.1.B03	189	0.41	-0.23	**0.34**
1.1.F02	201	0.32	-0.24	0.20
8.4.C08	201	0.42	-0.25	0.23
3.5.A04	228	0.58	0.23	**-0.32**

Table 5.5. Haplotypes derived from the same RFLP probe with sequence homology to beta-glucosidase

Marker	Linkage group	Marker frequency	Smut	Eldana
.....				
BGluD3	13	0.49	**-0.29**	0.19
BGluH5	13	0.45	-0.25	0.14
BGluH1	13	0.44	-0.23	0.10
.....				
5.1.B06	36	0.67	0.16	-0.13
8.4.B04	36	0.62	0.21	-0.15
BGluD1	36	0.57	**0.36**	**-0.26**
.....				
4.3.D08	118	0.49	0.08	0.15
8.1.C02	118	0.29	-0.06	**0.26**
BGluD2	118	0.23	0.10	0.13
.....				
5.1.C04	147	0.25	-0.21	0.12
BGluH4	147	0.18	**-0.34**	**0.28**
5.3.C01	147	0.31	-0.14	0.09
8.4.B03	147	0.22	**-0.39**	0.12
.....				

Four different linkage groups or haplotypes contain markers derived from the BGlu probe. For smut, two LGs contain markers for resistance, one has markers for susceptibility, and one LG is non-significant or has a neutral effect. These haplotypes may represent allelic variation across a homologous locus, or they may represent variation across distinct non-homologous loci coding for beta-glucosidases. DNA sequence information has demonstrated ancient genome duplication in many plant species including simple genomes such as *Arabidopsis* (The Arabidopsis Genome Initiative, 2000) as well as complex genomes such as cereals (e.g. Paterson *et al.*, 2004). Because of this, it is not possible to determine if these haplotypes are homologous or not at the coarse level of resolution inherent in the linkage disequilibrium map. For two of the linkage groups, *viz.* 36 and 147, the negative correlation between smut and eldana phenotype is apparent. One of these – 36 – is at relatively high frequency in the population. For breeding purposes it may be desirable to select individuals lacking this haplotype in order to reduce susceptibility to smut, even if an allele for eldana resistance is lost. This is because from Table 5.3 it is known that many other eldana resistance markers exist which may be positively selected for.

A more extensive set of haplotypes that may be homologous is illustrated in Table 5.6. These linkage groups all have markers derived from an RFLP probe homologous to Jacalin – a plant lectin implicated in plant defense responses (e.g. Chisholm *et al.*, 2000). Within this set, linkage group 3 contains an extensive region associated with smut resistance, and LG 208 is a short fragment associated with susceptibility. The remaining four LGs are not significantly associated with smut, although LG 139 has a marker moderately associated with resistance at r = -0.22. Linkage group 47 contains a marker significantly associated with eldana susceptibility, and two others moderately associated at r = 0.24 and r = 0.22. Linkage group 3 would be an obvious target for selection, as the smut resistance is not correlated with eldana susceptibility and the LG is at low frequency in the population. Making cross combinations between parents having this linkage group, e.g. N21, N17, 74M0659, N22 etc, would result in progeny populations where LG3 is present at high frequency. Within-family selection could then be used to select individual genotypes with suitable phenotypic value for other traits such as yield and sucrose content, and these could be used in turn as parents in a gene/allele pyramiding strategy.

Linkage group 47 may be desirable for selecting against in terms of eldana susceptibility, as it is not strongly correlated with smut resistance, and is a relatively high frequency in the population. In this case, rare genotypes without the susceptibility allele could be positively selected in order to decrease the frequency of the undesirable linkage group in the population. Although in this case alleles are not being pyramided, this strategy is essentially the inverse of a GAP strategy, and will be included in the objectives of allele pyramiding.

The observations made above on the potential to select for or against certain linkage groups have been made on an empirical basis, however. In order to utilize the map fully in a molecular breeding strategy for resistance to smut and/or eldana, a more thorough approach is required.

Table 5.6. Linkage groups containing markers derived from a Jacalin-like RFLP probe.

Marker	Linkage group	Marker frequency	Smut	Eldana
1.1.D02	3	0.21	**-0.40**	0.00
4.2.C03	3	0.23	**-0.37**	-0.01
JacH8	3	0.21	**-0.32**	-0.04
JacH3	3	0.21	**-0.32**	-0.04
6.7.D01	3	0.34	**-0.40**	0.08
6.6.D01	3	0.30	**-0.32**	0.06
3.7.A06	3	0.29	**-0.28**	0.11
6.5.E06	3	0.48	-0.11	-0.19
.....				
8.7.F01	27	0.69	0.08	-0.04
JacH4	27	0.73	0.10	-0.12
JacH7	27	0.70	0.08	-0.15
JacH1	27	0.69	0.08	-0.15
.....				
3.1.C03	47	0.67	0.05	0.11
3.4.C02	47	0.63	0.06	0.16
4.5.E03	47	0.64	-0.10	0.22
JacH5	47	0.64	-0.09	**0.27**
JacD6	47	0.64	0.03	0.18
JacD3	47	0.82	-0.07	0.24
NbsNo1D5	47	0.49	0.21	0.01
8.3.E04	47	0.65	0.07	0.01
.....				
3.6.A03	105	0.47	-0.08	0.06
JacH6	105	0.75	-0.16	0.02
JacD4	105	0.71	0.04	-0.12
.....				
8.3.C02	139	0.53	-0.22	0.04
JacD5	139	0.61	-0.09	0.17
JacH2	139	0.66	-0.01	-0.05
JacD2	139	0.83	0.11	0.00
.....				
JacD1	208	0.59	0.20	0.03
PrProD6	208	0.45	**0.34**	-0.10

5.3.2. Identification of marker ideotypes through stepwise regression

Stepwise multiple regression was performed in order to select combinations of six mapped markers associated with resistance to smut or to eldana, and the results are shown in Tables 5.7 and 5.8 respectively. In each case, markers used were restricted to those not showing a strong negative correlation between the two traits. The markers shown in Table 5.7 individually explain between 4% and 16% of the phenotypic variation in smut resistance, with the full model of six markers accounting for 54% of the variation. The individual effects of the markers are all significant at $P = 0.015$ or less. The full model is highly significant at $P < 0.0001$. The smut resistant ideotype has a predicted rating of 1.5, and does not result in a significant increase in susceptibility to eldana. Comparing this with Table 3.10b from Chapter 3.3.4 shows that this set of markers explains a similar amount of the variation in smut resistance, but has only two markers in common, *viz.* 1.1.D02 and 4.7.E03. The difference between the two sets of ideotypes lies in the fact that some markers from Table 3.10b were not mapped, and so were not included in the regression analysis.

Likewise, Table 5.8 shows six markers associated with resistance to eldana, with their corresponding effect on smut phenotype. The individual markers explain between 6% and 15% of the variation in eldana resistance, with the full regression model accounting for 59% of the variation. Although the eldana resistance ideotype results in a predicted smut rating of 6.5, this increase in susceptibility is not significant, with the full regression model for eldana having a F value of 0.51 and an associated probability of $P = 0.80$. Comparing against Table 3.11b from Chapter 3.3.4 shows that none of the markers are common, but the new resistance ideotype derived from the subset of 841 mapped markers ascribes a similar amount of variation in eldana resistance to the ideotype constructed using the full set of 1331 available markers.

A more detailed examination of the individual linkage maps for the markers associated with smut or eldana in the multiple regression models will give added insight into the use of markers and linkage disequilibrium maps in sugarcane breeding.

Table 5.7. Stepwise regression for markers associated with smut. The effect of the same markers on eldana is also shown. The predicted phenotypic score for the resistant and susceptible ideotype is given for both traits, as well as the R^2 values and significance levels for the individual markers, and the full regression model.

	SMUT		R^2	ELDANA		R^2
	Effect	Prob	marker	Effect	Prob	marker
Constant	7.572			4.843		
1.1.D02	-2.265	<0.0001	0.16	0.138	0.80	0.00
3.1.A02	-1.569	0.001	0.11	0.293	0.62	0.00
7.2.C02	-1.373	0.001	0.08	0.550	0.31	0.01
8.3.E02	1.417	0.004	0.04	0.116	0.86	0.00
4.7.E03	0.921	0.013	0.09	0.098	0.84	0.00
7.6.C07	-0.831	0.015	0.07	-0.058	0.90	0.00
R^2	0.54			0.02		
F-value	13.63	<0.0001		0.24	0.96	
Ideotype						
111001 (R)	1.5			5.8		
000110 (S)	9.9			5.1		

Table 5.8. Stepwise regression for markers associated with eldana. The effect of the same markers on smut is also shown. The predicted phenotypic score for the resistant and susceptible ideotype is given for both traits, as well as the R^2 values and significance levels for the individual markers, and the full regression model.

	ELDANA		R^2	SMUT		R^2
	Effect	Prob	marker	Effect	Prob	marker
Constant	6.361			5.337		
3.7.C01	-1.808	<0.0001	0.15	0.055	0.93	0.00
1.1.E01	-1.288	<0.0001	0.12	0.310	0.546	0.01
3.6.B07	1.146	<0.0001	0.14	-0.408	0.427	0.01
4.3.B05	1.173	0.001	0.06	-0.294	0.594	0.00
3.4.B03	-0.919	0.002	0.07	0.387	0.422	0.01
1.2.B02	-0.928	0.003	0.11	0.411	0.407	0.02
R^2	0.59			0.04		
F-value	16.17	<0.0001		0.51	0.80	
Ideotype						
110011 (R)	1.4			6.5		
001100 (S)	8.7			4.6		

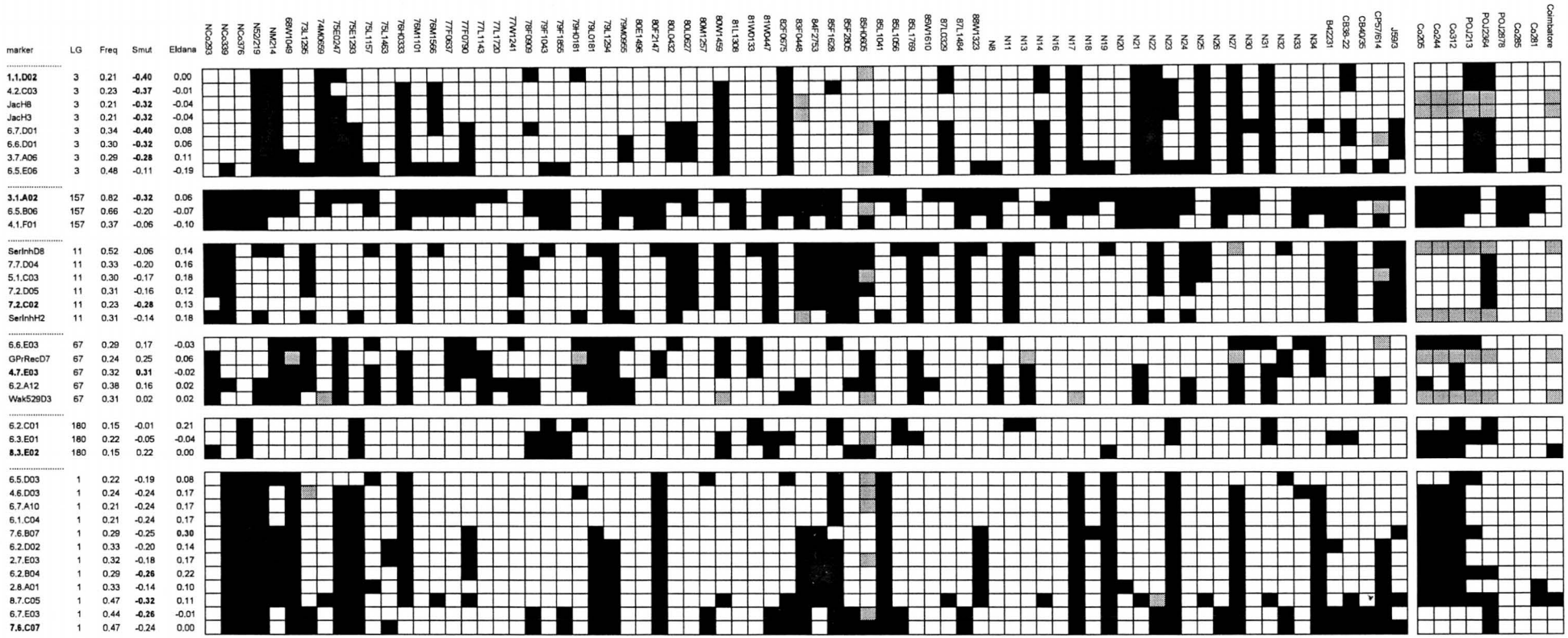
5.3.3. *Identification of important linkage groups associated with smut resistance or susceptibility, and tracing their ancestral origin.*

The map of the six linkage groups or haplotypes containing the markers associated with smut resistance as shown in Table 5.7 is given in Figure 5.3. The marker with the largest effect on smut rating, *viz.* 1.1.D02, is located on LG 3. This linkage group is at fairly low frequency in the population, and in terms of the ancestral genotypes included in this study, is found only in POJ213 and POJ2364. It is absent in POJ2878 and CO285, parents of the missing ancestor CO421, as well as CO312, CO244 and NCo376, so this LG did not enter the breeding population through the lineage giving rise to the "NCo" varieties, as depicted in Figure 5.1. Within the breeding population, LG 3 occurs in the genotype NM214, as well as genotypes containing NM214 in their pedigree viz. N21, N22, 82F0675, N27, N52/219 (Table 5.1). Genotype 75E0247 has the variety M168/32 as a great-grandparent, and M168/32 is a parent of NM214. The parents of M168/32 are POJ2878 and Uba Marot, (F1 hybrid between unknown *S. officinarum* and *S. spontaneum* clones), so Uba Marot is the most likely source or MRCA of LG3 in this lineage, as LG3 is absent form POJ2878.

A second origin for LG3 in the breeding population can be found in the lineage of CB38/22, N17 and N25. Although CB38/22 has the two Jacalin derived RFLP markers absent, this is likely to be a scoring error, as these markers are present in N17, an offspring of CB38/22. The *S. officinarum* clone D74 is a parent of CB38/22, and is also found in the lineage of N25, as it is a grandparent of Co419. It is likely that D74 is the origin of LG3 in this lineage, which may imply that Uba Marot inherited LG3 from its *S. officinarum* parent. For the other genotypes in the breeding population that have LG 3, (80W1459, 74M0659, N14, 76H0333, N31 and 87L0329) the complete lineage is either absent, or unknown due to the genotypes being derived from poly-crosses.

Linkage group 1 containing the marker 7.6.C07 is of particular interest. Although within the breeding population it tends to occur as a single linkage group, within the ancestral clones it seems to appear as two separate linkage groups – the first consisting of the first 10 markers (e.g. CO205, CO244 and CO312), and the second containing the two markers 6.7.E03 and 7.6.C07 (POJ2364). In the breeding population, LG1 occurs in NCo376 and NCo339, as well as offspring derived from these genotypes. LG1 appears to represent linkage disequilibrium due to population structure and not physical linkage. The two sub-groups of LG1 appear to have been inherited by NCo376 and NCo339 from CO312 and CO421, and have tended to be passed on together to their offspring by chance, giving the appearance of co-segregation. Looking at the frequency of the individual markers making up LG1 shows that the last two markers are at much higher frequency than the first nine markers. In addition, the first group

Figure 5.3. Haplotypes containing the markers associated with smut resistance ideotype from Table 5.7. A black square indicates the presence of the marker, while a white square indicates absence. A grey square signifies missing data. Markers significant at r > |0.25| are highlighted in bold.

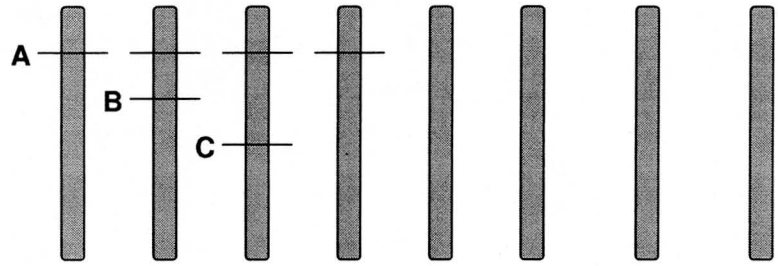


of markers has some association with eldana susceptibility, whereas the last two show no correlation with eldana. None of the genotypes within the mapping population not derived from the SASRI breeding programme, *viz.* CP57/614, CB40/35, J59/3, CB38/22 and NM214 show the presence of the composite LG1, further suggesting its origin due to population structure. LG1 should therefore be separated into two independent linkage groups. The second group containing the marker 7.6.C07 can be retained for use in breeding, as it is associated with smut resistance and is not correlated with eldana susceptibility. This perhaps illustrates a special case of the usefulness of an LD mapping approach to molecular breeding, despite the fact the apparent linkage is caused by population structure. Selecting individuals based on the presence of marker 7.6.C07 in the absence of LD information would result in the selection of genotypes such as 75E0247, N17, N23 *etc.*, that have this marker present, but also have the composite 'haplotype' containing markers associated with eldana susceptibility. The fact that population structure was detected through a coalescence interpretation of the map means that individuals segregating for the presence of only the desirable LD fragment, such as 81W0447, 73L1295, N22 *etc.* can be selected for breeding, to avoid increasing the frequency of the undesirable haplotype in the population. Disequilibrium due to structure can therefore be used to improve the efficiency of breeding, and will be referred to as structural disequilibrium (SD), to differentiate it from disequilibrium due to physical linkage (LD).

Linkage group 180 also appears unusual. Although it is present in CO312 and NCo376, it exists in a partial form in the other sugarcane ancestors, and the individual markers tend to segregate in the breeding population. In addition, the marker 8.3.E02 is associated with smut susceptibility and is uncorrelated with eldana rating, while the marker 6.2.C01 is uncorrelated with smut resistance, but associated with eldana susceptibility. One explanation for this pattern of segregation is that the apparent association is caused by population structure, in a similar manner to that illustrated by LG 1. There is an alternative explanation for the pattern of association displayed by the three markers in LG180, however. In a polyploid, a marker at high frequency may be linked to a marker at low frequency. This is illustrated in Figure 5.4, depicting a set of eight homologous chromosomes in an octaploid.

In Figure 5.4, marker A occurs on four homologous chromosomes, whereas markers B and C, occur on one homologue only, and are physically not linked. Because they are both linked to marker A, however, the branch and bound mapping algorithm used for mapping will assemble them into the same linkage group, placed on either side of marker A. LG 180 may represent this type of situation, with marker 6.3.E01 equivalent to marker A from Figure 5.4, and 6.2.C01 and 8.3.E02 equivalent to markers B and C respectively.

Figure 5.4. Illustration of hypothetical linkage arrangement of markers A, B and C in an octaploid, giving rise to false physical linkage. The vertical bars represent eight homologous chromosomes.



If this is indeed the case, markers 6.2.C01 and 8.3.E02 will represent allelic diversity on different linkage groups of the same homologous chromosome set, rather than a single linkage groups as depicted in Figure 5.3. To differentiate this class of allelic disequilibrium due to polyploidy from that due to linkage (LD) or population structure (SD), it will hereafter be referred to as homology disequilibrium (HD). It is necessary to note, however, that with this dataset it is not possible to differentiate between the different classes of disequilibrium, and any assignment of linkage groups in the map from Figure 5.2 is hypothetical.

In order to determine if the additional information on linkage disequilibrium, structure disequilibrium and homology disequilibrium can be used to improve the effectiveness of molecular breeding, the multiple regression shown in Table 5.7 was repeated, this time including two additional markers, one from LG 1 (7.6.B07), and one from LG 180 (6.2.C02). These markers are both located on the undesirable part of the composite LGs 1 and 180 representing putative HD and/or SD disequilibrium. From the regression model, the predicted rating for smut and eldana was calculated for the original smut resistance ideotype, *viz.* 111001, with either the presence or absence of the two additional markers from LGs 1 and 180. The results are shown in Table 5.9.

Comparing Table 5.8 with Table 5.7 shows that the results for smut are similar in both tables. The extended regression model of eight markers explains 58% of the variation in smut rating, with the two ideotypes consisting of the original resistance ideotype with either the presence or absence of the two new markers both predicting smut resistance. The case for eldana, however, is different. In the original regression model (Table 5.7), none of the individual markers had a significant effect on eldana rating, and the full model was non-significant, with the F-probability being $P = 0.96$.

Table 5.9. Stepwise regression using the six markers associated with smut resistance from Table 5.7, as well as the markers from putative SD linkage group 1 (7.6.B07) and PD linkage group 180 (6.2.C01) having an undesirable correlated effect with eldana. The ideotype given is for smut resistance, with either the presence or absence of the two additional markers.

	SMUT		**R^2**	**ELDANA**		**R^2**
	Effect	**Prob**	**marker**	**Effect**	**Prob**	**marker**
Constant	7.690			4.777		
1.1.D02	-2.307	<0.0001	0.16	0.199	0.702	0.00
3.1.A02	-1.485	0.001	0.11	0.014	0.981	0.00
7.2.C02	-1.493	<0.0001	0.08	0.760	0.137	0.01
8.3.E02	1.616	0.001	0.04	-0.134	0.830	0.00
4.7.E03	1.071	0.004	0.09	-0.211	0.659	0.00
7.6.C07	-0.585	0.100	0.07	-0.561	0.230	0.00
7.6.B07	-0.693	0.077	0.06	1.409	0.007	0.09
6.2.C01	-0.855	0.060	0.00	1.146	0.057	0.04
R^2	0.58			0.17		
F-value	11.54	<0.0001		1.74	0.10	
Ideotype						
11100100	1.8			5.2		
11100111	0.3			7.7		

In the extended model, however, the two new markers explain 9% and 4% of the variation in eldana rating respectively, at significance levels of $P = 0.007$ and $P = 0.057$. The full model of eight markers has an F-probability of $P = 0.10$, which although not considered statistically significant using the conventional arbitrary threshold of $P = 0.05$, is nevertheless suggestive that the two additional markers may have a real effect on eldana resistance. Comparing the two ideotypes shows that the presence of the additional markers results in a predicted eldana rating of 7.7, which is a substantial increase over the prediction of 5.2 when these markers are absent. In theory, therefore, the additional information gained on HD and/or SD from the map in Figure 5.3 can contribute to increasing in the effectiveness of breeding for smut resistance while not negatively influencing eldana susceptibility.

In practice, the real difference between using the six-marker model or the eight-marker model for selection will depend on the effect the two models have on the choice of desired parent combinations. Analysis of cross marker vectors as described in Chapter 3.3.5 was done for the six marker model from Table 5.7 and the eight-marker model from Table 5.9. From the 2926 possible cross combinations, those giving rise to the predicted progeny

ideotype for the original six-marker model, *viz.* 111001, (Table 5.7) were extracted. In addition, the combinations were restricted to those between genotypes not susceptible to smut, with phenotypic ratings less than 5. This resulted in a list of 42 cross combinations giving the desired smut resistance ideotype. Cross vector analysis as explained in Chapter 3.3.5 was carried out, to calculate the percentage of progeny expected to fall into progeny vector classes with predicted smut ratings of less than 3.5, and predicted eldana rating of less than 5.5.

For the eight marker model, the two extra markers, 7.6.B06 and 6.2.C01, were then added to the analysis. With the additional markers there are now six markers segregating, and $2^6 = 64$ progeny cross vectors possible. For each cross vector – i.e. marker combination, the predicted smut and eldana rating was calculated using the regression model from Table 5.9. For illustration, the 64 possible progeny vectors and their predicted smut and eldana ratings are given in Table 5.10. Individual progeny vectors resulting from segregation for the markers in the smut resistant cross ideotype range in predicted smut rating from 1.8 to 7.7, and in eldana rating from 4.2 to 7.5.

As explained in Chapter 3.3.5, each of the 42 cross combinations shown in Table 5.10 will segregate in the progeny, depending on the cross vector value for each marker – i.e. 0, 1 or 2. For each cross vector, the proportion of progeny expected in each of the 64 possible progeny vector categories shown in Table 5.10 was calculated. Thee percentage of progeny falling into vector classes predicted to be resistant to smut (rating less than 3.5), and not susceptible to eldana (rating less than 5.5) was then derived. This is shown in Table 5.11, along with the percentage of progeny falling into the same resistance ranges estimated from the six-marker model, as explained in the previous paragraph.

The difference between the results from the six- and eight- marker models lies in the fact that the extended model takes the structure and/or homology disequilibrium observed in LGs 1 and 180 into account. The six-marker model over-estimates the number of progeny expected to fall in the desired predicted phenotypic category, as it does not take the likely co-segregation of undesirable markers into account. For example, nearly a quarter of the progeny of the cross between N23 and CB38-22 are predicted to have desirable phenotype for smut and eldana based on the six-marker model. This cross, however, has the undesirable ideotype for the co-segregating markers on LGs 1 and 180. When this is taken into account by the eight-marker model, the cross is predicted to give only 7% of progeny with the desired phenotype.

Table 5.10. Possible progeny vectors for six segregating markers, with their predicted smut and eldana rating, derived from Table 9.

Marker vector								smut	eld
1	1	1	0	0	1	1	1	1.76	6.98
1	1	1	0	0	1	1	0	2.62	5.84
1	1	1	0	0	1	0	1	2.46	5.58
1	1	1	0	0	1	0	0	3.31	4.43
1	1	1	0	0	0	1	1	2.35	7.55
1	1	1	0	0	0	1	0	3.21	6.40
1	1	1	0	0	0	0	1	3.04	6.14
1	1	1	0	0	0	0	0	3.90	4.99
1	1	0	0	0	1	1	1	1.76	6.98
1	1	0	0	0	1	1	0	2.62	5.84
1	1	0	0	0	1	0	1	2.46	5.58
1	1	0	0	0	1	0	0	3.31	4.43
1	1	0	0	0	0	1	1	2.35	7.55
1	1	0	0	0	0	1	0	3.21	6.40
1	1	0	0	0	0	0	1	3.04	6.14
1	1	0	0	0	0	0	0	3.90	4.99
1	0	1	0	0	1	1	1	3.25	6.97
1	0	1	0	0	1	1	0	4.10	5.82
1	0	1	0	0	1	0	1	3.94	5.56
1	0	1	0	0	1	0	0	4.80	4.41
1	0	1	0	0	0	1	1	3.83	7.53
1	0	1	0	0	0	1	0	4.69	6.39
1	0	1	0	0	0	0	1	4.53	6.12
1	0	1	0	0	0	0	0	5.38	4.98
1	0	0	0	0	1	1	1	3.25	6.97
1	0	0	0	0	1	1	0	4.10	5.82
1	0	0	0	0	1	0	1	3.94	5.56
1	0	0	0	0	1	0	0	4.80	4.41
1	0	0	0	0	0	1	1	3.83	7.53
1	0	0	0	0	0	1	0	4.69	6.39
1	0	0	0	0	0	0	1	4.53	6.12
1	0	0	0	0	0	0	0	5.38	4.98
0	1	1	0	0	1	1	1	4.07	6.79
0	1	1	0	0	1	1	0	4.93	5.64
0	1	1	0	0	1	0	1	4.76	5.38
0	1	1	0	0	1	0	0	5.62	4.23
0	1	1	0	0	0	1	1	4.66	7.35
0	1	1	0	0	0	1	0	5.51	6.20
0	1	1	0	0	0	0	1	5.35	5.94
0	1	1	0	0	0	0	0	6.20	4.79
0	1	0	0	0	1	1	1	4.07	6.79
0	1	0	0	0	1	1	0	4.93	5.64
0	1	0	0	0	1	0	1	4.76	5.38
0	1	0	0	0	1	0	0	5.62	4.23
0	1	0	0	0	0	1	1	4.66	7.35
0	1	0	0	0	0	1	0	5.51	6.20
0	1	0	0	0	0	0	1	5.35	5.94
0	1	0	0	0	0	0	0	6.20	4.79
0	0	1	0	0	1	1	1	5.56	6.77
0	0	1	0	0	1	1	0	6.41	5.62
0	0	1	0	0	1	0	1	6.25	5.36
0	0	1	0	0	1	0	0	7.10	4.22
0	0	1	0	0	0	1	1	6.14	7.33
0	0	1	0	0	0	1	0	7.00	6.19
0	0	1	0	0	0	0	1	6.83	5.92
0	0	1	0	0	0	0	0	7.69	4.78
0	0	0	0	0	1	1	1	5.56	6.77
0	0	0	0	0	1	1	0	6.41	5.62
0	0	0	0	0	1	0	1	6.25	5.36
0	0	0	0	0	1	0	0	7.10	4.22
0	0	0	0	0	0	1	1	6.14	7.33
0	0	0	0	0	0	1	0	7.00	6.19
0	0	0	0	0	0	0	1	6.83	5.92
0	0	0	0	0	0	0	0	7.69	4.78

Table 5.11. Parental cross combinations resulting in progeny with predicted smut rating less than 3.5, and predicted eldana rating less than 5.5. The prediction is given for the 6-marker model (i.e. markers only) and the 8-marker model (i.e. markers and map information).

Parent 1	Parent 2	Cross marker vector								6-markers	8-markers
		M1	M2	M3	M4	M5	M6	M7	M8	% progeny with smut < 3.5 and eldana < 5.5	
CB38-22	82F675	2	2	1	0	0	2	0	0	33.8	42.2
CB38-22	N22	2	2	1	0	0	2	0	0	33.8	42.2
CB38-22	87L0329	2	2	1	0	0	1	0	0	27.1	28.1
CB38-22	N26	1	2	1	0	0	2	0	0	22.9	28.1
CB40/35	CB38-22	1	2	1	0	0	2	0	0	22.9	28.1
CB38-22	N52/219	2	2	1	0	0	2	1	0	33.8	21.1
CB38-22	80L0432	1	2	2	0	0	1	0	0	14.6	18.8
82F0675	80L0432	1	2	1	0	0	1	0	0	18.6	18.8
CB38-22	N24	1	2	1	0	0	1	0	0	18.6	18.8
CB38-22	N30	1	2	1	0	0	1	0	0	18.6	18.8
N22	80L0432	1	2	1	0	0	1	0	0	18.6	18.8
85F1628	CB38-22	1	2	2	0	0	2	0	1	17.5	14.1
CB38-22	83F0448	1	2	2	0	0	2	1	0	17.5	14.1
CB38-22	NCo339	1	2	2	0	0	2	1	0	17.5	14.1
J59/3	CB38-22	1	2	2	0	0	2	1	0	17.5	14.1
83F0448	82F0675	1	2	1	0	0	2	1	0	22.9	14.1
85F1628	82F0675	1	2	1	0	0	2	0	1	22.9	14.1
85F1628	N22	1	2	1	0	0	2	0	1	22.9	14.1
85L1056	CB38-22	1	2	1	0	0	2	0	1	22.9	14.1
J59/3	82F0675	1	2	1	0	0	2	1	0	22.9	14.1
J59/3	N22	1	2	1	0	0	2	1	0	22.9	14.1
N22	83F0448	1	2	1	0	0	2	1	0	22.9	14.1
N22	NCo339	1	2	1	0	0	2	1	0	22.9	14.1
NCo339	82F0675	1	2	1	0	0	2	1	0	22.9	14.1
CB38-22	84F2753	1	2	2	0	0	1	1	0	14.6	9.4
CB38-22	N11	1	2	2	0	0	1	0	1	14.6	9.4
80L0432	N52/219	1	2	1	0	0	1	1	0	18.6	9.4
82F0675	N11	1	2	1	0	0	1	0	1	18.6	9.4
83F0448	87L0329	1	2	1	0	0	1	1	0	18.6	9.4
84F2753	82F0675	1	2	1	0	0	1	1	0	18.6	9.4
85F1628	87L0329	1	2	1	0	0	1	0	1	18.6	9.4
J59/3	87L0329	1	2	1	0	0	1	1	0	18.6	9.4
N22	84F2753	1	2	1	0	0	1	1	0	18.6	9.4
N22	N11	1	2	1	0	0	1	0	1	18.6	9.4
NCo339	87L0329	1	2	1	0	0	1	1	0	18.6	9.4
83F0448	N52/219	1	2	1	0	0	2	2	0	22.9	7.0
85F1628	N52/219	1	2	1	0	0	2	1	1	22.9	7.0
J59/3	N52/219	1	2	1	0	0	2	2	0	22.9	7.0
N23	CB38-22	1	2	1	0	0	2	1	1	22.9	7.0
NCo339	N52/219	1	2	1	0	0	2	2	0	22.9	7.0
84F2753	N52/219	1	2	1	0	0	1	2	0	18.6	4.7
N52/219	N11	1	2	1	0	0	1	1	1	18.6	4.7

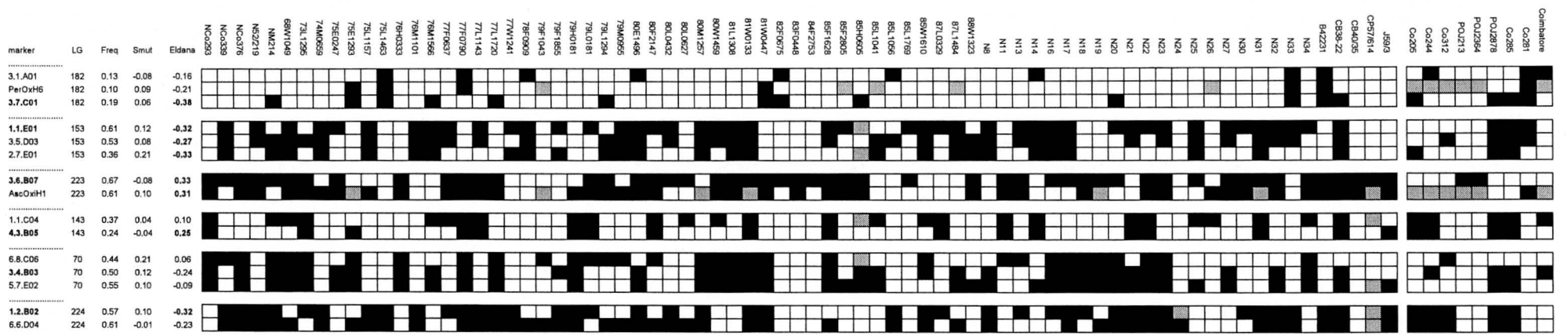
Information gained from the map has therefore resulted in increased precision of cross prediction. Those parent combinations resulting in less than 10% of progeny expected to have predicted smut rating less than 3.5 and eldana rating less than 5.5 will be lower priority when designing a crossing plan. Using the model derived from the map showing linkage disequilibrium, as well as structure and/or homology disequilibrium allowing crossing efforts to focus on the parent combinations shown at the top of Table 5.11, as these have a higher probability of producing progeny resistant to smut, without being susceptible to eldana. Using the regression model derived from markers only, and not taking of linkage and co-segregation into account, would have miss-directed effort to cross combinations with a low probability of producing desirable progeny.

5.3.4. *Identification of important linkage groups associated with resistance or susceptibility to eldana, and tracing their ancestral origin.*

A similar exercise was repeated for eldana. Mapped haplotypes containing the eldana resistance markers shown in Table 5.8 are given in Figure 5.5. These are all small fragments consisting of two or three markers only. The strongest resistance marker, 3.7.C01 is located on LG 182, which is at low frequency in the population. Only three individuals have this complete linkage group; B42231, 75L1463 and N33. From Table 5.2 it can be seen that B42231 is a parent of 75L1463, which in turn is a parent of N33. In the ancestral population, only CO281, which is a grandparent of B42231, has this group present, and is the likely source of these genes in the breeding population. As this haplotype is rare in the population, using genotypes such as N33 and 75L1463 in as parents in a GAP crossing strategy would be effective at increasing its frequency in the germplasm.

Linkage group 70, like LG 1, also appears to be a composite of two separate haplotypes displaying disequilibrium due to population structure, and not physical linkage. The first marker, 6.8.C06 is present in CO312, but not CO244 or POJ213. Its origin is likely to be Kansar, via CO213, which unfortunately is not present in the ancestral population. The two markers 3.4.B03 and 5.7.E02 are present in POJ2878 and CO285, so have likely been inherited via CO421. Both haplotypes occur in NCo376, and have tended to co-segregate in the progeny of NCo376 in the breeding programme. The haplotype with the markers 3.4.B03 and 5.7.E02 is desirable, as it is associated with eldana resistance. The haplotype with marker 6.8.C06 is undesirable, as it is not associated with eldana resistance, but with smut susceptibility. The fact that the two haplotypes have tended to co-segregate suggests that their combination may result in a beneficial effect, such as heterosis for other desirable traits, and their association has been maintained within the population by selection. Some

Figure 5.5. Haplotypes containing the markers associated with eldana resistance ideotype from Table 5.8. A black square indicates the presence of the marker, while a white square indicates absence. A grey square signifies missing data. Markers significant at r > |0.25| are highlighted in bold.



commercial varieties, however, such as N21, N8 and N32, lack the 6.8.C06 haplotype, while retaining the 3.4.B03/5.7.E02 portion. This suggests that through genetic recombination, the unknown portion of the 6.8.C06 haplotype involved in the beneficial effect can be dissociated from the 6.8.C06 marker region associated with smut susceptibility, and desirable recombinant individuals selected. The benefit of the LD map is that these individuals can be recovered by selecting for marker 3.4.B03 and against marker 6.8.C06.

Linkage group 153, is also strongly associated with resistance, but is at relatively high frequency in the population. Although it is present in its full form in POJ2878, and in some individuals in the mapping population, it also may result from disequilibrium due to structure, as markers 1.1.E01 and 3.5.D03 segregate from 2.7.E01 with fairly high frequency in the population. These two sub-groups are also present in the ancestor germplasm, represented by CO285 and CO205 respectively. Marker 2.7.E01 is moderately associated with smut susceptibility, while markers 1.1.E01 and 3.5.D03 show a lower degree of association. 2.7.E01 is also at lower frequency than the other two markers, indication that the association could also be due to homology, as illustrated in Figure 5.4. Within this apparently composite co-segregation groups, it may be beneficial selecting for the presence of marker 1.1.E01, and for the absence of 2.7.E01.

The two markers from LGs 153 and 70 were added to the original regression model shown in Table 5.8, to derive an eight-marker regression model shown in Table 5.12. The original six-marker eldana resistance ideotype is shown, with the two additional markers either present or absent. Comparing the two tables shows little difference in the results for eldana between the six-marker and eight-marker models. Marker 6.8.C06 is, however, significantly associated with smut susceptibility (P = 0.038), accounting for 4% of the variation in smut rating. The full regression model remains non-significant for smut rating, at P = 0.35, but the neutral effects of the other seven markers will overshadow the effect of 6.8.C06 in the full model. The ideotype with markers 2.7.E01 and 6.8.C06 present gives a predicted smut rating of 7.4, indicating that these individuals are likely to be more susceptible.

Table 5.12. Stepwise regression using the six markers associated with eldana resistance from Table 5.8, as well as the markers from putative SD linkage groups 153 (2.7.E03) and 70 (6.8.C06) having an undesirable correlated effect with smut. The ideotype given is for eldana resistance, with either the presence or absence of the two additional markers.

	ELDANA		R^2	SMUT		R^2
	Effect	**Prob**	**marker**	**Effect**	**Prob**	**marker**
CONSTANT	6.472			4.933		
3.7.C01	-1.816	<0.0001	0.15	0.138	0.822	0.00
1.1.E01	-1.227	0.001	0.12	0.311	0.566	0.01
3.6.B07	1.101	0.002	0.14	-0.509	0.352	0.01
4.3.B05	1.150	0.001	0.06	-0.338	0.537	0.00
3.4.B03	-0.823	0.017	0.07	-0.222	0.679	0.01
1.2.B02	-0.931	0.003	0.11	0.417	0.388	0.02
2.7.E01	-0.259	0.457	0.12	0.621	0.260	0.04
6.8.C06	-0.138	0.699	0.00	1.188	0.038	0.04
R^2	0.59			0.12		
F-value	11.99	<0.0001		1.14	0.35	
Ideotype						
11001*100*	1.7			5.6		
11001*111*	1.3			7.4		

In order to assess the likely impact of using the additional map information on molecular breeding decisions, cross vector analysis was done on this data set, as described for smut in 5.3.3 above. Details will not be repeated here. The results are shown in Table 5.13.

Twenty three parent cross vectors correspond to the original six-marker resistance ideotype of 110011. The six-marker model predicted low number of progeny with eldana rating of less than 3.5 and smut rating of less than 5.5. This is mainly due to the fact that the original model gave relatively high mean predicted smut rating of 6.5 (Table 5.8), and a low proportion of progeny segregating into maker vector classes with a predicted rating of less than 5.5. When the map information is taken into account, giving the eight-marker model shown in Table 5.12, some parent combinations result in a relatively high predicted frequency of progeny segregating into desirable marker and phenotypic classes. Priority can be given to making these crosses, and the number of progeny usually planted at Stage 1 of the selection programme can be increased, in order to increase the probability of selecting desirable phenotypes within these crosses. These crosses could also be targeted for using conventional marker-assisted selection, to identify the progeny with the specific desirable marker vectors. From Table 5.13 it is seen that none of the cross vectors correspond to the full eight-marker ideotype associated with reduced eldana susceptibility, as all combinations involve one or both parents with the undesirable marker 2.7.E01 from LG 153 (labeled M7 in

Table 5.13). Marker assisted selection could be used to identify superior progeny having the full eight-marker ideotype within appropriate crosses. These would then serve as parent germplasm in the next generation of GAP breeding to increase the frequency of desirable markers and linkage groups in the breeding population, while decreasing the frequency of undesirable markers or co-segregation groups.

Table 5.13. Parental cross combinations resulting in progeny with predicted eldana rating less than 3.5, and predicted smut rating less than 5.5. The prediction is given for the 6-marker model (i.e. markers only) and the 8-marker model (i.e. markers and map information).

Parent 1	Parent 2	Cross marker vector								6-marker	8-marker
		M1	M2	M3	M4	M5	M6	M7	M8	% progeny with eldana < 3.5 and smut < 5.5	
76M1566	N33	2	2	0	0	1	1	1	0	4.1	18.8
75L1463	76M1566	2	1	0	0	1	1	1	0	6.3	15.6
Co281	N8	1	2	0	0	1	1	1	0	3.1	15.6
Co281	79F1855	1	2	0	0	1	1	1	0	3.1	15.6
87L1484	N33	1	2	0	0	1	1	1	0	3.1	15.6
N8	N33	1	2	0	0	1	2	1	0	1.8	14.1
N33	79F1855	1	2	0	0	1	2	1	0	1.8	14.1
75L1463	87L1484	1	1	0	0	1	1	1	0	4.7	12.5
75L1463	N8	1	1	0	0	1	2	1	0	2.7	12.5
75L1463	79F1855	1	1	0	0	1	2	1	0	2.7	12.5
76M1566	N8	1	2	0	0	2	1	2	0	1.8	9.4
76M1566	79F1855	1	2	0	0	2	1	2	0	1.8	9.4
75L1157	76M1566	1	2	0	0	1	1	2	0	3.1	7.8
76M1566	74M659	1	2	0	0	1	1	2	0	3.1	7.8
76M1566	76M1101	1	2	0	0	1	1	2	0	3.1	7.8
Co281	85W1610	1	2	0	0	1	1	1	1	3.1	7.8
Co281	80M1257	1	2	0	0	1	1	1	1	3.1	7.8
85W1610	N33	1	2	0	0	1	2	1	1	1.8	7.0
80M1257	N33	1	2	0	0	1	2	1	1	1.8	7.0
75L1463	85W1610	1	1	0	0	1	2	1	1	2.7	6.3
75L1463	80M1257	1	1	0	0	1	2	1	1	2.7	6.3
76M1566	85W1610	1	2	0	0	2	1	2	1	1.8	4.7
76M1566	80M1257	1	2	0	0	2	1	2	1	1.8	4.7

5.4. Discussion

The use of the methodology developed in Chapter 4 was able to identify 231 co-segregating marker-groups within a population of 77 sugarcane genotypes used as parents in the SASRI breeding programme. These linkage groups appear to represent distinct classes of association due to disequilibrium, *viz*, that due to physical linkage (LD) that due to population structure (SD) and that due to homology (HD). Due to the limitations of the data set and the lack of a reference genetic map against which the LD map could be compared, assignment of linkage groups to LD, SD or HD classes was hypothetical.

In the case of LD due to physical linkage, the linkage groups represent haplotypes present in the ancestral *Saccharum* germplasm and maintained within modern germplasm due to a limited number of generations from the most recent common ancestors. By including ancestral clones in the population that was marker-typed, some haplotypes of interest could be traced back to their origin. Hypothetically, this would allow the re-creation of the early generation germplasm tailored for specific haplotype structure. For example, both desirable and undesirable haplotypes have been inherited into the breeding population through CO421 and CO312, the parents of the 'NCo' varieties. By re-creating the bi-parental crosses from which they were selected, viz, POJ2878 x CO285 and CO213 x CO244, within-family selection could be done for individuals with desirable phenotype (vigour etc) combined with the presence of desirable marker haplotypes and the absence of undesirable haplotypes (e.g. those associated with disease susceptibility). The selected individuals, full-sibs to CO421 and CO312, could then be inter-crossed to give an alternative 'NCo' population having an increased frequency of 'good' linkage groups, and a reduced frequency of 'bad' haplotypes. This is possible in sugarcane, as the crop is clonally propagated and ancestral clones are maintained in germplasm collections for exploitation.

Interpreting the population map within a coalescence framework to trace the ancestral origin of haplotypes identified cases of disequilibrium that are likely to be caused by population structure or homology effects. Analysis of the complete set of marker data as described in Chapter 3.3.2 did not detect any significant structure within the marker discovery and mapping population. This result is presumably because the large number of markers used obscured the cases of specific haplotypes co-segregating in specific lineages. It is well known that population structure can result in the false detection of marker-trait associations (e.g. Lander and Schork, 1994). This is because an individual's lineage will affect its probability of having any particular allele that varies across lineages, as well as the probability of having a particular phenotype. Any allele that shares a joint distribution with phenotype will therefore appear to be associated, but this association may be spurious, in terms of the allele being cause of the phenotype. (Rosenberg and Nordborg, 2006). The risk of using a spurious marker-trait association in breeding occurs if it is used for selection outside the lineage in which it was detected. In this situation, there may be no association at all of marker and phenotype. The risk is much smaller if the marker is used for selection within the lineage in which it was detected. In this case, although the marker may not be causally associated with phenotype, the fact that it co-segregates with phenotype within the pedigree implies that there is an increased probability of both marker and trait being transmitted to progeny together.

In the work reported here, disequilibrium due to structure or homology as evidenced in linkage groups 1 and 180 from Figure 5.3.was used to improve the prediction of expected progeny resistance to smut, as shown in Tables 5.11. This improvement will only be realised if the estimation of marker-trait effect used in the models is real and not spurious. The fact that in LG1 the two co-segregating haplotypes had different marker-trait associations with eldana rating implied that the association is not spurious, as if the association had been due to joint frequency distribution of phenotype with random alleles, their joint association with eldana would be expected to be equivalent. The same argument applies to the use of LG 180 in the molecular model used to predict smut resistance ideotype. The particular segregation pattern of markers in LG 180, and their association with smut and eldana suggests that this co-segregation group results from HD, and represents allelic variation within a Homology Group. This should perhaps be regarded as a special case of structure disequilibrium, as the disequilibrium results from co-segregation of homologous chromosomes carrying different alleles within certain lineages. The fact that the molecular breeding strategy advocated here relies on the use of the marker discovery population as the parents for breeding, followed my molecular characterisation of progeny from specific crosses means that the validity of the marker-trait associations can be tested in the next generation. Podlich *et al.* (2004), coined the term 'Mapping as You Go" (MAYG) to describe a strategy of continually revising marker estimates by remapping elite germplasm over cycle of selection. This philosophy is inherent within the GAP strategy proposed here.

Although disequilibrium *not* due to physical linkage is problematic within a fine mapping/map based cloning context, knowledge of haplotypes co-segregation due to structure is useful in a breeding context, as demonstrated in the improvement of cross prediction for smut and eldana resistance. In describing the genetic study of populations, Rosenberg and Nordborg (2002) remark that "analysis of polymorphism data must take the historical nature of the data into account", as "polymorphism data reflects a unique, complex, non-repeatable evolutionary history". One of the benefits of the mapping approach described here is that population structure for certain haplotypes is revealed through a coalescent interpretation. Population structure can then be used in a positive manner in terms of tracing the transmission of desirable or undesirable haplotypes through the lineage, instead of regarding structure as a 'nuisance' parameter in a statistical analysis. In fact it could be argued that in the study described in this chapter, the detection of structure was the main benefit derived from mapping. The predicted increase in breeding efficiency illustrated in Tables 5.11 and 5.13 is solely derived by accounting for co-segregation due to structure. If this had not been done, a six-marker model using mapped markers only would not have been any more efficient than a model using markers before the construction of the map.

One of the motivations behind this study was to identify allelic variation contributing to phenotype across homologous groups. The low number of sequence-based RFLP markers, however, meant that linkage groups identified could not be assigned to defined Homology Groups with any degree of confidence. Comparing AFLP gels from this study with gels from the R570 mapping population proved to be difficult, so common markers scored in both populations could not be identified. The LD map could therefore not be related to the sugarcane reference map of variety R570. An additional issue in sugarcane mapping is the number of markers required. For a genome comprising of ~100 chromosomes, ~4000 useable, low frequency markers would be required to place 20 markers per chromosome arm. If one then considers that in association mapping the haplotypes of several ancestor genotypes are being mapped, this number must be multiplied by the number of ancestors contributing to the genome pool. The number of marker required to provide coverage of the potential haplotype variation present within a breeding population is therefore very large. Current sequence based markers such as RFLP and SSR are not able to generate large numbers of polymorphisms. High throughput systems are therefore required, but anonymous markers such as the AFLPs used in this study have the disadvantage that they cannot be used to identify Homology Groups. In addition, the frequency of AFLP scoring errors of ~4% detected in this study is higher than desirable. In order to address these issues, a collaborative project funded by the International Consortium for Sugarcane Biotechnology is currently underway to map 80 genotypes from the SASRI breeding population, 10 ancestral clones and the reference variety R570 using Diversity Arrays Technology (DArT). DArT is essentially an array of sequence characterised fragments allowing a chip-based high throughput and automated technique for scoring polymorphic markers (Jaccoud *et al.*, 2001, Wenzl *et al.*, 2004). The sugarcane DArT array is designed for the scoring of ~4000 polymorphisms (Kilian, personal communication[1]), and will allow direct comparison between the LD map and the R570 map. This will facilitate the interpretation of the LD map, and enable issues such as the extent of structure disequilibrium and putative homology disequilibrium within the population to be investigated.

To be used routinely in breeding, molecular markers need to offer some advantage over the use of phenotype data alone in terms of making selection and crossing decisions. Comparing the efficiency of conventional breeding to molecular breeding is a complex issue, dependent on factors such as trait heritability, population size, selection intensity, the number of markers required, the relative cost of collecting phenotypic versus molecular data etc.

[1] Andrzej Kilian. Director: Diversity Arrays Technology Pty Ltd. email: a.kilian@diversityarrays.com

Studies have shown that marker assisted selection can be more efficient that phenotypic selection when trait heritability is low (e.g. Lande and Thompson, 1990, Knapp, 1998), trait phenotype is difficult or costly to measure (Yousef and Juvik, 2001, Xie and Xu, 1998) when linkage between marker and QTL is high (Dudley, 1993, Knapp, 1998) and in earlier generations of selection (Hospital *et al.*, 1997). Ironically, however, the statistical power of detecting marker-trait association is low when trait heritability is low (Moreau *et al.*, 1998). Individual estimates of efficiency are dependent on breeding strategy, population size, marker system, *etc*, and are difficult to compare objectively. None of the studies referenced here have considered negatively correlated traits, such as smut and eldana resistance. At this stage it is not possible to objectively compare conventional breeding versus molecular breeding for smut and eldana resistance. This will only be possible once the effectiveness of the GAP strategy has been validated. This will be done by phenotyping progeny from crosses with different predicted ratings for smut and eldana, in order to determine if the observed phenotype correlates with that predicted. No matter what the outcome of the validation exercise is, there is no doubt that the population-level map described here has provided valuable insights into the genetic structure of the breeding population, the contribution of ancestral clones to the current genetic diversity, and potential methodologies of exploiting this information in breeding strategies. Although it may be only the first step in developing a structured approach to the molecular breeding of sugarcane at SASRI, this work represents a significant advancement in the use of linkage disequilibrium or association mapping in sugarcane, and perhaps in crop plants in general. No similar report describing mapping *de novo* from association data has been reported in the literature to date for any crop species. The International Consortium for Sugarcane Biotechnology DArT maps are scheduled for completion towards the end of 2007. These will provide additional information on the genetic structure of sugarcane breeding germplasm, as serve as a validation study for the approach and methodology described here.

5.5. References

Ballhorn, DJ, Heil, M and Lieberei, R. 2006. Phenotypic plasticity of cyanogenesis in lima bean *Phaseolus lunatus* — activity and activation of β-glucosidase. Journal of Chemical Ecology 32: 261-275.

Breseghello, F and Sorrels, ME. 2005. Association mapping of kernel size and milling quality in wheat. Genetics 172: 1165-1177.

Chisholm, ST, Mahajan, SK, Whitham, SA, Yamamoto, ML and Carrington, JC. 2000. Cloning of the Arabidopsis RTM1 gene, which controls restriction of long-distance movement of tobacco etch virus. Proceeding of the National Academy of Sciences USA 97: 489-494.

Dudley, JW. 1993. Molecular markers in plant improvement: Manipulation of genes affecting quantitative traits. Crop Science 33: 660-668.

Goldstein, DB and Weale, ME. 2001. Population genomics: Linkage disequilibrium holds the key. Current Biology 11:R576-R579.

Hospital, F, Moreau, L, Lacoudre, F, Charcosset, A and Gallais, A. 1997. More on the efficiency of marker-assisted selection. Theoretical and Applied Genetics 95: 1181-1189.

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

Kingman, JFC. 1982a. On the genealogy of large populations. Journal of Applied Probability 19A; 27-43.

Kingman, JFC. 1982b. The coalescent. Stochastic Processes and their Applications. 13: 235-248.

Kingman, JFC. 2000. Origins of the coalescent: 1974-1982. Genetics 156: 1461-1463.

Knapp, S. 1998. Marker-assisted selection as a strategy for increasing the probability of selecting superior genotypes. Crop Science 38: 1164-1174.

Kraakman, ATW, Niks, RE, Van den Berg, PMMM, Stam, P and Van Eeuwijk, FA. 2004. Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars. Genetics 168: 435-446.

Kraakman, ATW, Martinez, F, Mussiraliev, B, van Eeuwijk, FA and Niks, RE. 2006. Linkage disequilibrium mapping of morphological, resistance and other agronomically relevant traits in modern spring barley cultivars. Molecular Breeding 17: 41-58.

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

Lander, ES and Schork, NJ. 1994. Genetic dissection of complex traits. Science 265: 2037-2048.

Moreau, L, Charcosset, A, Hospital, F and Gallais, A. 1998. Marker-assisted selection efficiency in populations of finite size. Genetics 148: 1353-1365.

Nahrstedt, A. 1985. Cyanogenic compounds as protecting agents for organisms. Plant Systematics and Evolution 150: 35-47.

Nagylaki, T. 1989. Gustav Malécot and the transition from classical to modern population genetics. Genetics 122: 253-268.

Nordborg, M. 2000. Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. Genetics 154: 923-929.

Palmer, LJ and Cardon, LR. 2005. Shaking the tree: mapping complex disease genes with linkage disequilibrium. The Lancet 366: 1223-1234.

Paterson, AH, Bowers, JE and Chapman, BA. 2004. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. Proceeding of the National Academy of Sciences USA 101: 9903-9908.

Podlich, DW, Winkler, CR and Cooper, M. 2004. Mapping as you go: an effective approach for marker-assisted selection of complex traits. Crop Science 44: 1560-1571.

Reich, DE, Cargill, M, Bolk, S, Ireland, J, Sabeti, PC, Richter, DJ, Lavery, T, Kouyoumjian, R, Farhadian, SF, Ward, R and Lander, ES. 2001. Linkage disequilibrium in the human genome. Nature 411: 199-204.

Rosenberg, NA and Nordborg, M. 2002. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nature Reviews Genetics 3: 380-390.

Rosenberg, NA and Nordborg, M. 2006. A general population-genetic model for the production by population structure of spurious genotype-phenotype associations in discrete, admixed or spatially distributed populations. Genetics 173: 1665-1678.

Simko, I, Costanzo, S, Haynes, K, Christ, BJ and Jones, RW. 2004. Linkage disequilibrium mapping of a *Verticillium dahliae* resistance quantitative trait locus in tetraploid potato (*Solanum tuberosum*) through a candidate gene approach. Theoretical and Applied Genetics 108: 217-224.

The *Arabidopsis* Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408: 796-815.

TropGENE database. http://tropgenedb.cirad.fr

Weir, BS. 1996. Genetic data analysis II. Sinauer Associates Inc, Sunderland, Massachusetts. Pp445.

Wenzl, P, Carling, J, Kudrna, D, Jaccoud, D, Huttner, E, Kleinhofs, A and Kilian, A. 2004. Diversity Arrays Technology (DArT) for whole-genome profiling of barley. Proceeding of the National Academy of Sciences USA 101: 9915-9920.

Xie, C and Xu, S. 1998. Efficiency of multistage marker-assisted selection in the improvement of multiple quantitative traits. Heredity 80: 489-498.

Yousef, GG and Juvik, JA. 2001. Comparison of phenotypic and marker-assisted selection for quantitative traits in sweet corn. Crop Science 41: 645-655.

Yu, J and Buckler, ES. 2006. Genetic association mapping and genome organization of maize. Current Opinion in Biotechnology 17: 155-160.

Zhang, W., Collins, A., Maniatis, N, Tapper, W and Morton, N.E. 2002. Properties of linkage disequilibrium (LD) maps. Proceeding of the National Academy of Sciences USA 99: 17004-17007.

5.6. APPENDIX A

/* LDMap; GAUSS programme for calculating marker-marker associations and ordering markers into linkage groups*/

```
format /rdt 5,4;
fname1 = "c:\\gauss50\\myprog\\alldata050204.xls";   /*data input file*/
ddata = xlsreadm(fname1,"B1:ca1336",1,"");  /*range for marker data*/
ndata = xlsreadm(fname1,"A1:A1336",1,"");  /*range for marker names*/
gdata = ddata';
ndata = ndata';
nrows = rows(gdata);
ncols = cols(gdata);
start = 2;  /* first row of data*/
nloci = ncols-start+1;
compcrit = 73; /* # of comparisons - i.e. without missing data*/
fq1critu = 0.85;
fq1critl = 0.10;
fq2critu = 0.85;
fq2critl = 0.10;
Rcrit = 0.5; /* to divide into high and low freq groups*/
r2crit1 = 0.48; /* r threshold for declaring disequilibrium */
rqcrit1 = 3;

print "Freq1 between  " fq1critl " and " fq1critu;
print "Freq2 between  " fq2critl " and " fq2critu;
print "Association > " r2crit1;
print "R/Q < " rqcrit1;
let mask[1,10] = 0 0 1 1 1 1 1 1 1 1;
let fmt[10,3] =
"-*.*s" 10 8
"-*.*s" 10 8
"-*.*lf" 5 0
"-*.*lf" 5 0
"-*.*lf" 6 2
"-*.*lf" 6 2
"-*.*lf" 6 2
```

```
"-*.*lf" 6 2
"-*.*lf" 6 2
"-*.*lf" 3 0;
let mask2[1,2] = 1 0;
let fmt2[2,3] =
"-*.*lf" 10 0
"-*.*s" 10 8;

print nloci;
firstc = start+1;
ncomp = (nloci*(nloci-1))/2;
print ncomp;
compmat = zeros(ncomp,10);
rcount = 1;
for c1(start,ncols,1);
   for c2(firstc,ncols,1);
      if c1 >= c2;
         continue;
      endif;
      n1 = ndata[1,c1];
      n2 = ndata[1,c2];
      x1 = gdata[.,c1];
      x2 = gdata[.,c2];
      xmat = x1~x2;
      scrit = (xmat[.,1] .== 0 .or xmat[.,1] .== 1) .and (xmat[.,2] .== 0 .or xmat[.,2] .== 1);
      xmat2 = selif(xmat,scrit);
      rxmat2 = rows(xmat2);
      if scalmiss(xmat2);
        continue;
      endif;
      sumxmat = sumc(xmat2);
      freq1 = sumxmat[1,1]/rxmat2;
      freq2 = sumxmat[2,1]/rxmat2;
      sumxmat2 = sortc(sumc(xmat2),1);
      Q = sumxmat2[1,1]/rxmat2;
      R = sumxmat2[2,1]/rxmat2;
      QR = Q * (1-R);
```

```
        cmat = vcx(xmat2);
        Dhat = cmat[1,2];
        rho = Dhat/QR;
        zz = corrx(xmat2);
        r2 = zz[1,2];
        compmat[rcount,1] = n1;
        compmat[rcount,2] = n2;
        compmat[rcount,3] = c1;
        compmat[rcount,4] = c2;
        compmat[rcount,5] = R; /* to get highest freq*/
        compmat[rcount,6] = r2;
        compmat[rcount,7] = freq1;
        compmat[rcount,8] = freq2;
        compmat[rcount,9] = R/Q;
        compmat[rcount,10] = rxmat2;
        rcount = rcount+1;
    endfor;
endfor;
compmat2 = rev(sortc(compmat,6));
ccrit1 = compmat2[.,6] .> r2crit1 .and
        compmat2[.,9] .< rqcrit1 .and
        compmat2[.,10] .> compcrit;

assoc1 = selif(compmat2,ccrit1);
assoc = rev(sortc(assoc1,6));
z = rows(assoc);
print;
d3 = printfm(assoc,mask,fmt);
print;
print "XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX";
print;
worklist = union(assoc[.,3],assoc[.,4],1);
pnum = rows(worklist);
permlist = (pnum*(pnum-1))/2;

create f1 = hold with x,2,8;
for a(1,pnum-1,1);
```

```
   for b(a+1,pnum,1);
      h1 = worklist[a,1];
      h2 = worklist[b,1];
      hm = h1~h2;
      hm1 = writer(f1,hm);
   endfor;
endfor;
open f2 = hold for update;
hm2 = readr(f2,permlist);
f1 = close(f1);
f2 = close(f2);

create fa = hold2 with x,10,8;
for aa(1,permlist,1);
   hh1 = hm2[aa,1];
   hh2 = hm2[aa,2];
   hcrit = compmat2[.,3] .== hh1 .and compmat2[.,4] .== hh2;
   workmat = selif(compmat2,hcrit);
   workmat1 = writer(fa,workmat);
endfor;
open fb = hold2 for update;
workmat2 = readr(fb,permlist);
fa = close(fa);
fb = close(fb);

zx = rows(assoc);
do while zx > 1;
   s1 = assoc[1,3];
   s2 = assoc[1,4];
   lcrit = assoc[.,3] .== s1 .or assoc[.,4] .== s1 .or assoc[.,3] .== s2 .or assoc[.,4] .== s2;
   g1 = selif(assoc,lcrit);
   rowzg = rows(g1);
   if rows(g1) ==1;
      lg1 = g1[1,3]~g1[1,1]|g1[1,4]~g1[1,2];
      print ".....";
      pg1 = printfm(lg1,mask2,fmt2);
      print ".....";
```

```
   g2 = g1;
   goto g1next;
endif;
g2 = g1;
list3 = union(g2[.,3],g2[.,4],1);
c3 = rows(list3);
comb3 = (c3 * (c3-1))/2;
u3count=1;
create f1g = holdg with x,2,8;
for a(1,c3-1,1);
   for b(a+1,c3,1);
      hg1 = list3[a,1];
      hg2 = list3[b,1];
      hgm = hg1~hg2;
       hgm1 = writer(f1g,hgm);
   endfor;
endfor;
open f2g = holdg for update;
hgm2 = readr(f2g,comb3);
f1g = close(f1g);
f2g = close(f2g);

dcount = 0;
create fag = hold2g with x,10,8;
for aa(1,comb3,1);
   hhg1 = hgm2[aa,1];
   hhg2 = hgm2[aa,2];
   hgcrit = assoc[.,3] .== hhg1 .and assoc[.,4] .== hhg2;
   if hgcrit == 0;
      dcount = dcount+1;
      goto hgout;
   endif;
   gx = selif(assoc,hgcrit);
   gx1 = writer(fag,gx);
   hgout:
endfor;
open fbg = hold2g for update;
```

```
g2 = readr(fbg,comb3-dcount);
fag = close(fag);
fbg = close(fbg);

{x,nxhold} = mapprocr(ndata,start,workmat2,g2);  /*Calls branch&bound subroutine*/
print "...............................";
nxp = printfm(x'~nxhold',mask2,fmt2);
print "...............................";
g1next:
gmat = g2[.,3]~g2[.,4];

if rows(gmat) == 1;
   gd1 = gmat[1,1];
   gd2 = gmat[1,2];
   dacrit = assoc[.,3] .== gd1 .and assoc[.,4] .== gd2;
   dgcrit = workmat2[.,3] .== gd1 .and workmat2[.,4] .== gd2;
   assocd = delif(assoc,dacrit);
   assoc = assocd;
   workmat2d = delif(workmat2,dgcrit);
   workmat2 = workmat2d;
   goto nextg1next;
endif;

rowg = rows(gmat);
dellist = union(g2[.,3],g2[.,4],1);
delnum = rows(dellist);

for aa(1,delnum,1);
   dd1 = dellist[aa,1];
   acrit = assoc[.,3] .== dd1 .or assoc[.,4] .== dd1;
   if acrit == 0;
      goto anext;
   endif;
   assoca = delif(assoc,acrit);
   assoc = assoca;
   anext:
   dcrit = workmat2[.,3] .== dd1 .or workmat2[.,4] .== dd1;
```

```
      workmatd = delif(workmat2,dcrit);
      workmat2 = workmatd;
   endfor;
nextg1next:

print;
zx = rows(assoc);
endo;
endall:

lgx = assoc[1,3]~assoc[1,1]|assoc[1,4]~assoc[1,2];
pgx = printfm(lgx,mask2,fmt2);
dxx = printfm(assoc,mask,fmt);
end;

/* mapproc2.src   Procedure for assembling marker order using branch and bound*/
proc(2) = mapprocr(ndata,start,workmat2,g2);
local ulist,rowu,ncomb,u2,ufreq,uc1,uc2,ucrit,usel,uloop,ucount,
nmark,tm2,x,nextx,x2,sx,myx,z,seq1,seq2,seq3,myy,
cnum,hsize,hmat,hcount,lcrit,tm3,mhold,mcrit,mmat,
mx,mhold2,bestc,bestp,nx,nxhold,nxidx;

format /rdt 3,2;
ulist = union(g2[.,3],g2[.,4],1);
rowu = rows(ulist);
ncomb = (rowu * (rowu-1))/2;
u2 = zeros(ncomb,3);
ufreq = zeros(ncomb,2);
ucount=1;
for a(1,rowu-1,1);
   for b(a+1,rowu,1);
      uc1 = ulist[a,1];
      uc2 = ulist[b,1];
      u2[ucount,1] = uc1;
      u2[ucount,2] = uc2;
      ucrit = workmat2[.,3] .== uc1 .and workmat2[.,4] .== uc2;
      usel = selif(workmat2,ucrit);
```

```
        if scalmiss(usel);
            goto uloop;
        endif;
        u2[ucount,3] = usel[1,6];
        ufreq[ucount,1] = usel[1,7];
        ufreq[ucount,2] = usel[1,8];
        uloop:
        ucount=ucount+1;
    endfor;
endfor;
/*print u2~ufreq;*/
print;
nmark = rows(ulist);

tm2 = u2|u2[.,2]~u2[.,1]~u2[.,3];

ulist = ulist';
format /rdt 3,1;
x = zeros(1,2);
x[1,1] = ulist[1,1];
x[1,2] = ulist[1,2];

for h(3,nmark,1);
    nextx = ulist[1,h];
    x2 = x~nextx;
    sx = cols(x)+1;
    myx = zeros(sx,sx);

    /* matrix of column coordinates */
    for i(1,sx,1);
        z = sx-i;
        if z == 0;
            continue;
        endif;
        seq1 = seqa(i,1,z)';
        seq2 = seqa(1,1,i)';
        seq3 = seq2~seq1;
```

```
    myx[i,.] = seq3;
    myx[i,i] = sx;
endfor;
myx[sx,.] = seqa(1,1,sx)';

/* matrix of marker number permutations */
myy = zeros(sx,sx);
for j(1,sx,1);
    for k1(1,sx,1);
        cnum = myx[j,k1];
        myy[j,k1] = x2[1,cnum];
    endfor;
endfor;

/* matrix */
hsize = sx*(sx-1);
hmat = zeros(hsize,4);
hcount = 1;
for k(1,sx,1);
    for l(1,sx-1,1);
        lcrit = ((tm2[.,1] .== myy[k,l]) .and (tm2[.,2] .== myy[k,l+1]));
        tm3 = selif(tm2,lcrit)~k;
        hmat[hcount,.] = tm3;
        hcount = hcount+1;
    endfor;
endfor;

mhold = zeros(sx,2);
for m(1,sx,1);
    mcrit = hmat[.,4] .== m;
    mmat = selif(hmat,mcrit);
    mx = sumc(mmat);
    mhold[m,.] = mx[3,1]~m;
endfor;
mhold2 = rev(sortc(mhold,1));

bestc = mhold2[1,2];
```

```
    bestp = myx[bestc,.];
    x = myy[bestc,.];
    nx = cols(x);
    nxhold = zeros(1,nx);
    for d(1,nx,1);
        nxidx = x[1,d];
        nxhold[1,d] = ndata[1,nxidx];
    endfor;
endfor;
retp(x,nxhold);
endp;
```

CHAPTER 6

CONCLUDING REMARKS

Sugarcane breeding is often referred to as a 'numbers game'. The nature of the crop precludes the use of inbred or isogenic lines, which means that single genes of interest cannot be selectively introduced into a stable, desirable genetic background through backcrossing. In addition, the polyploidy genome results in extreme linkage drag, as any particular genetic region desired from a selected parent in a cross-combination will be inherited along with four or five other copies on homologous chromosomes. Commercial varieties are required to be phenotypically stable under vegetative propagation over large areas for many years for a large number of commercially important traits. The result of these factors is that breeding programmes need to test large numbers of candidate genotypes over many years in order to improve the chances of identifying superior varieties. Any information that can be used to improve the efficiency of this process is valuable.

The work reported here provides new information and new tools to the sugarcane breeder that can be directly applied in variety development. Through association analysis within germplasm representing the broader breeding population, molecular markers correlated with resistance or susceptibility to smut and eldana have been identified. Using association between markers and traits within diverse germplasm resulted in the identification of a wider range of putative quantitative trait alleles than that present within a conventional mapping population.

The fact that the two traits of interest addressed in this study are phenotypically negatively correlated makes the application of markers in breeding particularly significant. Multi-marker regression models for resistance to each trait could be derived, taking the correlated effect of the markers on the second trait into account. By accounting for the negative correlation within the model, selection for resistance for one trait will not result in an increase in susceptibility in the second trait. 'Resistance', however, is not a single trait but can be the result of a number of different unrelated causes, many of which are not understood. For example, stalk fibre content is an effective resistance mechanism against the eldana stem borer, but is an undesirable economic trait for sugarcane. Stalk silicon content has also been implicated in resistance, and genetic differences in silicon accumulation between genotypes may play a part in varietal resistance. In addition to these physical mechanisms that may also include traits such as rind hardness, antibiosis can also contribute towards defence against insect herbivores, and it is likely that there are several or many mechanisms involved

in plant resistance. In the data set used in this study, the mechanism contributing to phenotypic resistance in unknown, and significant markers are likely to be associated with genes involved in a range of different mechanisms.

A statistical model comprising of a small number of markers is thus unlikely to account for a large percentage of the phenotypic variation observed in a diverse pool of germplasm. Specific models of different markers may, however, be effective at explaining the phenotype of subsets of genotypes sharing similar types of resistance. Phenotypic data on the forms of resistance that are easily measurable - such as fibre content or silicon content - could assist with this, as it may be possible to assign markers associated with phenotypic resistance to specific resistance mechanisms. This is something to consider for the future.

Work in association analysis in other crops reviewed in Chapter 2.6 has focussed on identifying marker-trait associations for use in breeding. The work reported here is unique in that in addition to identifying marker-trait associations, marker-marker associations have been used to identify haplotypes in disequilibrium. This has allowed the construction of a population-level disequilibrium map. This has not been reported in the literature for other crops to date. Validation of the LD mapping method using data collected at CIRAD detected several cases of physical linkage disequilibrium extending more than 50 cM, with one haplotype corresponding to 120 cM of the genetic map of the cultivar R570. This is considerably more than that reported for other crops (Chapter 2.6), and reflects the particular genome and breeding history of sugarcane. Mapping whole populations as described here requires the existence of extensive disequilibrium. For this reason it is likely that this approach will not be appropriate in crops where disequilibrium extends over short distances, such as maize. The approach may be useful, however, in species where there has been a strong founder effect in recent breeding history, such as cacao and oil-palm.

The population-level map derived for the SASRI germplasm, coupled with coalescence interpretation using the genealogies of the mapped individuals, has provided valuable insight into the genetic structure of the population. The likely origin of desirable or undesirable haplotypes in linkage disequilibrium could be traced back to some ancestral genotypes. This offers the possibility of re-creating early-generation hybrid germplasm with specific haplotype structure. 'Haplotypes' which appear to result from co-segregation of two different linkage groups were also identified. The cause of the co-segregation is likely to be population structure. Accepted methods for identifying population structure conducted before map construction had not detected the presence of structure within the data. Coalescence interpretation of the map, however, revealed cases of disequilibrium due to co-segregating

markers within lineages. The structure identified could be used to improve the regression models for resistance to smut and eldana, as shown in Chapters 5.3.3 and 5.3.4. Ironically, it appears that detecting structure for certain co-segregating groups is a valuable application of whole-population mapping. It is known that population structure can result in the detection of false marker-trait associations. As only a very small subset of markers/linkage groups from the entire map are used in subsequent breeding, it is possible to investigate each candidate marker for possible false association before choosing to use it. The presence of potential false association – either with phenotype or with other markers – within linkage groups that will not be used for decision-making is not important in the context of using selected markers for molecular breeding.

The use of markers in breeding for traits such as resistance to smut and eldana remains a complex issue. If a trait phenotype is influenced by several mechanisms, each controlled by several loci, with each locus having several possible alleles, the number of markers required to explain the phenotype will be large. Three major issues then come into play. First of all, the cost of marker-typing a population will increase as the number of markers increases, and the issue of cost effectiveness will arise. Secondly, from a statistical perspective a regression model of many markers derived from a population of moderate size (e.g. 100-200 genotypes) is likely to be over-fitted. This is because the number of possible marker-classes (2^x, where x = number of segregating markers) is likely to be much greater than the number of genotypes in the population. Thirdly, as more markers are included in the breeding model, the number of genotypes or cross combinations matching the desired ideotype decreases. Issues around genetic variation in later generations of breeding then become important. For these reasons, in this study the number of segregating markers in the GAP breeding model was restricted to six, with a total of eight markers for marker-typing (for both smut and eldana, two markers for susceptibility in the model were absent in the desired ideotype). It is not possible to claim, however, that this is an optimum model.

Despite the complexity and uncertainty, this work represents an important advance in the molecular breeding of sugarcane at SASRI, with a clear path forward in application and further development. Using the cross prediction methodology described in Chapter 5.3.3 and 5.3.4, combinations expected to be resistant and susceptible to smut and eldana based on marker ideotype have be made. These are being screened at the family level for phenotypic reaction to the two traits using existing SASRI protocols. Comparing the observed and predicted phenotype will provide a validation for the individual markers and cross prediction approach, without having to marker-type all the progeny populations. Within some crosses of interest, resistant and susceptible individuals will be sampled in order to validate the

prediction at the level of the genotype – i.e. resistant versus susceptible ideotypes segregating within a single family. These data will allow an estimation of the efficiency as well as the cost effectiveness of molecular breeding using a GAP strategy in sugarcane.

This study was based on a pre-existing population for which some molecular data was available. The obvious question is; could the study have been improved by designing a population *de novo* specifically for LD mapping, and what effect could this have had on the results? The issues involved, which are somewhat inter-related, are population structure, population size and the number (and type) of markers. The predicted extent of linkage disequilibrium within a population can be manipulated by considering the genealogies of the individuals within the population. For example, if individuals were chosen based on the fact they all shared a common grandparent, LD would be more extensive in haplotypes inherited from that ancestor. It would also have the effect, however, of narrowing the genetic base of the population, and so limit the potential to identify additional QTAs present in more diverse germplasm. Choice of population type to maximise LD or maximise genetic variation would depend on the specific objectives. In this study, the population was sufficiently diverse to identify range of useful markers from a variety of different ancestors; if the population had been chosen to maximize LD, it is likely that some of these would not have been identified.

In terms of population size, increasing the number of individuals would increase the precision of LD estimates between individual marker-pairs, and large population sizes would be required for fine-scale mapping. Doubling the population size of this study to 150 individuals would most probably have resulted in some minor changes in marker order on identified haplotypes, and also most probably identified some additional haplotypes or marker-trait associations involving weak LD at the threshold of the statistical significance level. The question is, however, would these differences have resulted in an additional benefit in terms of information gained or application in breeding to justify the extra time and cost of genotyping more individuals? I would speculate that genotyping extra individuals would not have resulted in a substantial difference to the results presented. In addition, increasing the population size also increases the likelihood of introducing population sub-structure, which may result in an increase in the false association detection rate. I would suggest that a population size of 80-100 individuals is a reasonable compromise between the time and cost involved in genotyping, the precision of association estimation and managing population structure. As many automated genotypic platforms are based on a 96-well or 96-capilliary format, this is the logical population size to use (minus any control samples that may be run) for any study using these technologies.

If additional resource were available for genotyping, increasing the number of markers on a specified population would be more advantageous than increasing the population size. Increasing the density of markers across the genome increases the chances of detecting smaller regions of LD, and also increases the chance of detecting rare polymorphisms. In addition, the denser the map, the greater the likelihood of being able to identify which Homology Group a particular linkage group or haplotype belongs to. This facilitates identifying potential allelic relationships between identified marker-trait associations. Type of marker may also play a role, but perhaps has been overstated. Sequence specific markers such as SSRs and RFLP are generally regarded as locus specific, allowing the detection of allelic relationships. Information from genome sequences (*Arabidopsis* and maize, for example) has shown, however, that extensive gene duplication occurs across the genome. This calls into question a strict allelic interpretation of polymorphic bands revealed by sequence specific probes and primers. The only way to reliably confirm the existence of allelic relationships is through comparison to a reference map.

These considerations have been taken into account in new work being done by SASRI through an International Consortium for Sugarcane Biotechnology collaboration. A population of 80 individuals and 10 ancestors is being genotyped on a DArT array (96-well format) designed to detect ~4000 polymorphisms. This population already has data on 1956 AFLP markers. The combined AFLP/DArT data-set of more than 5000 markers thus represents a substantial increase over that used in the study reported here, and should result in an LD map covering larger portions of the genome. At the same time, a population 100 progeny derived from self-fertilization of R570 is being genotyped using the same DArT array. Data from the R570 progeny will be used to extend the current reference map, substantially increasing the marker density and facilitating the assignment of linkage groups to Homology Groups. Comparing the new LD map against the extended reference map of R570 will allow the extent and patterns of LD within the population, and allelic relations between haplotypes, to be investigated at a much greater resolution than that achieved to date. In addition phenotypic data has been collected for a range of yield component, sucrose, pest and disease, and agronomic traits. Examining the number, distribution and strength of marker-trait associations for complex traits such as yield and sucrose content will allow the potential to use a GAP molecular breeding strategy for complex traits to be evaluated. The results of this work currently underway will be of great value in determining the future direction and strategy of sugarcane molecular breeding at SASRI and in other sugarcane breeding programmes in general.