Abstract

LIU, KEJUN. Software and Methods for Analyzing Molecular Genetic Marker Data (under the direction of DR SPENCER V. MUSE)

Genetic analysis of molecular markers has allowed biologists to ask a wide variety of questions. This dissertation explores some aspects of the statistical and computational issues used in the genetic marker data analysis. Chapter 1 gives an introduction to genetic marker data, as well as a brief description to each chapter. Chapter 2 presents the different genetic analyses performed on a large data set and discusses the use of microsatellites to describe the maize germplasm and to improve maize germplasm maintenance. Considerable attention is focused on how the maize germplasm is organized and genetic variation is distributed. A novel maximum likelihood method is developed to estimate the historical contributions for maize inbred lines. Chapter 3 covers a new method for optimal selection of a core set of lines from a large germplasm collection. The simulated annealing algorithm for choosing an optimal k-subset is described and evaluated using the maize germplasm as an example; general constraints are incorporated in the algorithm, and the efficiency of the algorithms is compared to existing methods. Chapter 4 covers a two-stage strategy to partition a chromosomal region into blocks with extensive within-block linkage disequilibrium, and to select the optimal subset of SNPs that essentially captures the haplotype variation within a block. Population simulations suggest that the recursive bisection algorithm for block partitioning is generally reliable for recombination hotspots identification. Maximal entropy theory is applied to choose optimal subset of SNPs. The procedures are evaluated analytically as well as by

simulation. The final chapter covers a new software package for genetic marker data analysis. The methods implemented in the package are listed. A brief tutorial is included to illustrate the features of the package. Chapter 5 also describes a new method for estimating population specific F-statistics and an extended algorithm for estimating haplotype frequencies.

SOFTWARE AND METHODS FOR ANALYZING MOLECULAR GENETIC MARKER DATA

by
KEJUN LIU

A dissertation submitted to the Graduate Faculty of

North Carolina State University

in partial fulfillment of the

requirements for the Degree of

Doctor of Philosophy

BIOINFORMATICS

Raleigh

2003

APPROVED BY:

SPENCER V. MUSE CHAIR OF ADVISORY COMMITTEE	BRUCE S. WEIR
EDWARD BUCKLER	MONTSERRAT FUENTES

To my wife, my parents and my parents-in-law

Biography

Kejun Liu (Jack) was born in Shaodong, Hunan Province, China on June 18, 1978. In 1992, he was enrolled in the special class for gifted youth (SCGY) in Wuhan University, Hubei, China, where he received a Bachelor of Science degree in Molecular Biology in 1997. After graduation, Kejun was employed as a database programmer by Founder Cooperation, an IT company in China. After one year he was appointed as the chief architect of the Department of System Integration in Wuhan Founder, a subdivision of the Founder Cooperation. Kejun entered the department of Statistics in North Carolina State University in 1999. During his stay in the Biomathematics program, he worked as a research assistant under the instruction of Dr. Spencer V. Muse. Since August 2000, Kejun has been studying for his PH.D in the Bioinformatics Research Center (BRC), North Carolina State University, and received the Genomic Science Fellowship from the program of Genomic Science. He was a member in the maize evolutionary genomics project, providing support in statistics methodology and software development. He also performed independent research under the direction of his advisor Dr. Spencer V. Muse.

Acknowledgements

I would like to express my deepest gratitude to Dr. Spencer V. Muse for being my mentor and guide throughout my graduate studies. His insightful instructions will never go unremembered.

I would also like to thank the members of my advisory committee, Dr. Bruce S. Weir, Dr. Edward Buckler and Dr. Montserrat Fuentes, for their helpful advice and encouragement. They have been a wonderful committee and great resources of valuable ideas which I have found incredibly helpful.

I would like to express my appreciation to all of the members in the maize evolutionary genomics project for their advice and stimulating scientific discussions, especially Dr. John Doebley and Dr. Major Goodman who have been an enormous help to me in both maize genetics and maize evolutionary studies. Special thanks to all of the faculty and staff of the Bioinformatics Research Center. I also thank Doug Robinson, Xiang Yu and Jieye Yu for their friendship and helpful comments.

I thank Li Li for her wonderful support.

Table of contents

List of tables

List of figures

Chapter 1

Introduction

GENETIC MARKER DATA

The advent of genetic marker technology designed to detect naturally occurring polymorphisms at the DNA level has become an invaluable and revolutionizing tool for both applied and basic diagnostic studies of plant, animal and human genomes as well as for microorganisms. A genetic marker is an identifiable physical location on a chromosome whose inheritance can be monitored. Genetic markers are locus specific and polymorphic in the studied populations. The various established and widely used genetic marker techniques include approaches for detecting restriction fragment length polymorphism (RFLP), randomly amplified polymorphic DNA (RAPD), amplified fragment length polymorphism (AFLP), single nucleotide polymorphism (SNP) and Mini- and Microsatellites (or simple sequence repeats, SSRs).

Clearly, strategies based on the polymerase chain reaction (PCR) genotyping techniques have the highest potential for routine diagnosis. Along this line, SSRs and SNPs (Removal of the PCR step is now possible for scoring SNPs) have become the predominant genetic markers. Microsatellite Markers (SSRs) are highly polymorphic, easily genotyped and occur evenly throughout the genome, making them especially suitable for genetic analysis. Single nucleotide polymorphisms (SNPs) are single base pair positions in genomic DNA at which the variant sequence type has a frequency of at least 1% in the sample. SNPs, essentially biallelic polymorphisms, have lower mutation rates than do repeat sequences such as SSRs, and therefore they are not as informative as microsatellite markers. SNPs are, however, more frequent than SSRs, providing markers near to or in the locus of interest, some located within the gene coding or regulatory

region, which can directly influence protein structure or expression levels, giving insights into disease mechanisms. Approximately 90% of human DNA polymorphism, which accounts for a larger fraction of observed differences between individuals, is due to SNPs (Patil *et al.* 2001).

Fundamental and applied population genetics, quantitative genetics and forensic genetics depend heavily on the availability of genetic markers. Markers are used by population geneticists to investigate origin, genetic diversity, gene flow, population structure and relationships among species; by evolutionist to describe genetic relationships among species between different genera; by geneticists to construct full coverage or QTL maps; and by forensic scientists to identify individuals. Markers are now widely used in the search for genes affecting human diseases, based on associations between genetic markers and the traits of interest.

MAIZE INBREDS DATA ANALYSIS

For the past 4 years I have been involved in the maize evolutionary genomics project. Directed by John Doebley at the University of Wisconsin, this project involves 5 different campuses and 20+ scientists. The project has produced an enormous amount of molecular marker data. The maize inbreds data I have analyzed includes 100 microsatellites from 340 assays representing 260 worldwide maize inbred lines. I applied both traditional and new methods to address a variety of questions. The heart of the analysis is to infer the genetic structure of the inbred lines. I applied traditional clustering algorithms, phylogenetic approaches, as well as a model-based clustering approach to get an integrated genetic structure and substructure. This genetic structure, along with the genetic diversity it revealed, provides a useful reference to plant breeders and maize geneticists. Another objective of the project was to infer the historical sources for different inbred groups as well as for single accessions. I developed a new maximum likelihood approach to study the admixture model. The model provides a statistical framework for defining appropriate parameters, developing estimate of these parameters, and comparing the statistical properties of estimates. The estimates were largely consistent with known pedigree information. A variety of other analyses were performed, including investigations of linkage disequilibrium, of relationships of inbreds to open-pollinated sources, and of relationships of genetic data to known pedigrees. This work is summarized in chapter 2. This project also motivated me to develop the Optimal K-Subset theory and to implement the PowerMarker software package (Chapter 3 and Chapter 5).

CHOOSING CORE SET OF LINES BY MAXIMIZING ALLELIC RICHNESS

Maize geneticists often ask questions of the sort: "How do I select a small number of lines to represent the total diversity found in a larger germplasm collection?" Previous methods for answering this question were not based on genetic data, and therefore had a number of undesirable properties. In addition to estimating the population structure of maize inbreds to provide a hierarchical approach for this problem, I have also worked on a general framework to study this type of question. The framework provides an abstract model to maximize a single criterion or multiple criteria of interest by effectively searching through the solution space, allowing for different combinatorial algorithms and network design models to be used. For the maize CoreSet the criteria could be the total allelic richness or genetic diversity based on available marker data. This problem was proven to be NP-Complete (Garey 1979), as the possible number of subsets increases exponentially when the sample size increases linearly. The major approach I used to search the global solution space is based on simulated annealing. Chapter 3 essentially solved the searching question, and extends general simulated annealing to constrained cases. Constrained maximization is of special interest in practice. For example, how do I select 50 lines from a set of 1000, with the constraints that the 50 lines include at least 2 representatives of each major landrace and at least 1 line from each major geographic region? Our work provided an unprecedented approach for satisfying customized constraints.

BLOCK PARTITIONING AND HAPLOTYPE TAGGING

Recent studies suggested the human chromosome appeared to be organized as haplotype blocks. Within these blocks, high linkage disequilibrium (LD) and limited haplotype variation were observed (Patil *et al.* 2001; Gabriel *et al.* 2002). The data of haplotype blocks have left several uncertainties concerning the exact block definition and block boundaries. Existing methods for block partitioning are either dependent on haplotype data (Patil *et al.* 2001) or specific to their applications (Gabriel *et al.* 2002). In chapter 4, a general algorithm for block partitioning is proposed to maximize the possibility of recombination hotspot identification based on pairwise linkage Disequilibrium (LD). In our method, a block can be defined as a chromosomal region where recombination hotspots do not exist, and the correlation between pairwise LD and physical proximity is not significant. The properties of the algorithm were studied using population simulations.

Within each block, only a relatively small number of SNPs (referred as tagging SNPs) is required to retain most of the haplotype variation (Clayton 2001). One intention of choosing tagging SNPs is to reduce the high genotyping cost. More importantly, association studies based on tagging SNPs will be much more useful than existing single marker or sliding-window based methods, as a positive association between a trait and tagging SNPs will localize the candidate gene to the block level, whereas a negative result will rule out the whole block region. In chapter 4, a formal relationship between tagging SNPs and haplotype entropy is given, maximal entropy theory is applied to obtain the minimal set of tagging SNPs, and the algorithms are evaluated using population simulations.

POWERMARKER – A SOFTWARE FOR GENETIC DATA ANALYSIS

PowerMarker (http://www.powermarker.net) is an Integrated Analysis Environment (IAE) for genetic marker data. The objective of the software is to provide a comprehensive set of data analysis and data management tools for population genetics researchers. PowerMarker was written in C# to take full advantage of the Microsoft .Net Framework and future platform independence. PowerMarker builds a powerful user interface around both new and traditional statistical methods for population genetic analysis, including summary statistics, population structure, linkage disequilibrium, association tests, coalescence simulation, haplotype estimation, phylogeny, and all the new methods developed in this thesis. PowerMarker is designed to interact seamlessly with Excel, and interacts with TreeView and other population genetic software such as ARLEQUIN and STRUCTURE. We have observed that PowerMarker routines are up to 50 times faster than those in other packages.

In the development of PowerMarker, several existing methods were extended at the methodology level and/or implementation level. For example, Population specific F-statistics (Weir and Cockerham 1984; Weir and Hill, 2002) were extended by taking population-specific inbreeding coefficients into account, and therefore are more appropriate for populations not in Hardy-Weinberg equilibrium. At the implementation level, an extremely efficient haplotype estimation procedure was implemented with multiple optimizations of the traditional E-M algorithm (Slatkin 1995). We also extend the algorithm to a more general case by allowing for pedigree incorporation. Chapter 5

covers a brief tutorial to PowerMarker, and gives an introduction to the methods implemented in the package.

REFERENCES

Clayton D, Choosing a set of haplotype tagging SNPs from a larger set of diallelic loci. ftp-gene.cimr.cam.ac.uk/software (2001).

Excoffier L, Slatkin M, Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* **12**: 921-927 (1995).

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D, The structure of haplotype blocks in the human genome. *Science* **226**: 225-2229 (2002).

Garey MR and Johnson DS, In: Computers and Intractability (Freeman, New York), 222 (1979).

Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, Nguyen BT, Norris MC, Sheehan JB, Shen N, Stern D, Stokowski RP, Thomas DJ, Trulson MO, Vyas KR, Frazer KA, Fodor SP, Cox DR, Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**: 1719-1723 (2001).

Weir BS and Hill WG. Estimating F-Statistics. *Annu. Rev. Genetic.* **36**, 721-750 (2002)

Weir BS and Cockerham CC, Estimating F-statistics for the analysis of population structure. *Evolution* **38**: 1358-1370 (1984)

Chapter 2

Genetic Structure and Diversity among Maize Inbred Lines as Inferred from DNA Microsatellites

Kejun Liu[1], Major Goodman[2], Spencer Muse[1], J. Stephen Smith[3], Ed Buckler[4], and John Doebley[5]

[1]Department of Statistics, North Carolina State University, Raleigh, NC 27695,

[2]Department of Crop Science, North Carolina State University, Raleigh, NC 27695,

[3]Crop Genetics Research and Development, DuPont Agriculture and Nutrition, Pioneer Hi-Bred International, Johnston, IA 50131,

[4]USDA-ARS and Department of Genetics, North Carolina State University, Raleigh, NC 27695,

[5]Laboratory of Genetics, University of Wisconsin, Madison, WI 53706

Running Title: Diversity among Maize Inbreds

Key words: maize, *Zea*, SSR, microsatellite, inbreds

Corresponding Author:

John Doebley

445 Henry Mall

Madison, Wisconsin 53705

Phone: 608 265 5803

Fax: 608 262 2976

jdoebley@facstaff.wisc.edu

ABSTRACT

Two hundred sixty maize inbred lines, representative of the genetic diversity among essentially all public lines of importance to temperate breeding and many important tropical and subtropical lines, were assayed for polymorphism at 94 microsatellite loci. The 2039 alleles identified served as raw data for estimating genetic structure and diversity. A model-based clustering analysis placed the inbred lines in five clusters that correspond to major breeding groups plus a set of lines showing evidence of mixed origins. A "phylogenetic" tree was constructed to further assess the genetic structure of maize inbreds, showing good agreement with the pedigree information and the cluster analysis. Tropical and subtropical inbreds possess a greater number of alleles and greater gene diversity than their temperate counterparts. The temperate Stiff Stalk lines are on average the most divergent from all other inbred groups. Comparison of diversity in equivalent samples of inbreds and open-pollinated landraces revealed that maize inbreds capture less than 80% of the alleles in the landraces, suggesting that landraces can provide additional genetic diversity for maize breeding. The contributions of four different segments of the landrace gene pool to each inbred group's gene pool was estimated using a novel likelihood-based model. The estimates are largely consistent with known histories of the inbreds and indicate that tropical highland germplasm is poorly represented in maize inbreds. Core sets of inbreds that capture maximal allelic richness were defined. These or similar core sets can be used for a variety of genetic applications in maize.

INTRODUCTION

Maize (*Zea mays* L. ssp. *mays*) inbred lines represent a fundamental resource for studies in maize genetics and breeding. While maize inbreds have been used extensively in hybrid corn productions (Anderson and Brown 1952; Troyer, 2001), they have also been critical for diverse genetic studies including the development of linkage maps (Burr *et al.* 1988), quantitative trait locus mapping (Edwards *et al.* 1987; Austin *et al*. 2001), molecular evolution (Henry and Damerval 1997; Ching *et al.* 2002), developmental genetics (Poethig 1988; Fowler and Freeling 1996), and physiological genetics (Crosbie *et al.* 1978). Most recently, a set of diverse maize inbreds has been employed to perform the first phenotype-genotype association analyses in a plant species (Thornsberry *et al.* 2002) and estimate linkage disequilibrium in maize (Remington *et al.* 2002; Tenaillon *et al.* 2002).

The intelligent exploitation of maize inbreds for genetic analyses requires a detailed knowledge of genetic and historical relationships among these lines and an understanding of the partitioning of genetic diversity among them. For example, developmental mutants of maize can exhibit strikingly difference phenotypes when assayed in the genetic background of different maize inbred lines (Poethig 1988). A knowledge of the relationships among lines would help identify a set of inbreds that have maximal diversity of the analysis of the effects of genetic background. Single Nucleotide Polymorphism (SNP) discovery in maize can be optimized by selecting a set of lines that capture the maximum number of alleles or haplotypes. Use of maize inbreds in

association analyses requires that population structure among lines be factored into the analysis (Thornsberry *et al.* 2002).

In this paper, we analyze the genetic structure and diversity among maize inbred lines using DNA microsatellites or simple sequence repeats (SSR) and a comprehensive set of 260 inbreds that well represent the diversity available among current and historic used lines. We show that these lines can be partitioned in three major groups, that diversity is greatest among tropical inbreds, that maize inbreds capture about 80% of the allelic diversity in open-pollinated lines, and that one-level of population structure can not fully explain linkage disequilibrium among inbreds. We also define core sets of inbreds that capture maximal allelic diversity for given sample sizes, investigate the relationship between pedigree and genetic distance, and identify the portions of the broader maize germplasm pool from which maize inbreds were derived.

MATERIAL AND METHODS

Plant Materials

A set of 260 inbred lines representing a sample of the most important public lines from the US, Europe, Canada, South Africa, and Thailand, along with lines from the International Center for the Improvement of Maize and Wheat (CIMMYT) and the International Institute of Tropical Agriculture (IITA) were chosen to represent the diversity available among current and historic lines used in breeding. These include essentially all public lines of importance to temperate breeding and many of the important tropical and subtropical lines. The 260 lines and their pedigrees are listed in Supplemental Table S1. Seed of most lines are still available from their original sources (see http://statgen.ncsu.edu/panzea/), but we have also provided seed samples to both the North Central Regional Plant Introduction Station (NCRPIS) at Ames, IA and to the National Seed Storage Laboratory at Ft. Collins, CO. Most, if not all, lines should be available from the NCRPIS in 2004.

SSR Genotyping

The lines were genotyped at Celera AgGen (Davis, CA, USA). The details of the genotyping have been published elsewhere (Romero-Severson *et al.* 2001). Briefly, DNA was extracted from individual plants by the cTAB method, and the microsatellite regions were amplified by PCR with florescent-labeled primers, PCR products were size-separated on Applied Biosystems fragment analyzers equipped with GeneScan software, and the PCR products were classified to specific alleles (bins) by GeneScan and Genotyper software programs (Romero-Severson *et al.* 2001). We used 100 SSR loci

that are evenly distributed throughout the genome. A list of the SSR loci with their chromosomal locations has been deposited as Supplemental Table S2. Primer sequences are available at the MaizeGDB (http://www.maizegdb.org).

Preanalysis

We began with 264 lines some of which were assayed two to four times for the 100 SSR loci giving a total of 339 assays. Of the 33,900 SSR-genotypes, 4.3% amplified more than one band per inbred line, perhaps because of residual heterozygosity, contamination, or the amplification of similar sequences in two separate genomic regions. In order to minimize the effect of contamination, we dropped seven assays with heterozygosity > 0.20, an unexpectedly high value for maize inbreds. Further, four other assays, which represented the sole assays for four lines, were excluded from the study because their position in a preliminary cluster analysis was strongly discordant with their known pedigrees, suggesting a seed or sample mix-up. We also dropped four loci with mean within-line heterozygosity > 0.10, suggesting that these loci did not faithfully amplify a single locus or that allele-calling was problematic. We dropped two loci with availability (defined as 1 – proportion of missing or null data) < 0.80, suggesting that the locus could not be amplified in the PCR reaction for many lines. The final data set consists of 260 lines and 94 loci.

We performed multiple SSR assays for some lines. So that each inbred is represented only by a single entry in our data set for statistical analyses, we built consensus genotypes for inbreds that were assayed more than once. The main criterion for constructing the

consensus genotype was that any allele with frequency > 25% is counted, but if there are three or more alleles that have frequency > 25% then we regard the genotype is missing. The second criterion is that if one assay gave a null phenotype but the other a viable allele, then the inbred was considered homozygous for the visible allele. Since there was a high degree of concordance among assays, inferred consensus genotypes based on these criteria represent only 1.9% of the final data set.

Summary statistics and tests

We used PowerMarker (Liu and Muse 2003) to calculate observed heterozygosity, gene diversity (or expected heterozygosity), number of private alleles, number of group-specific alleles, pairwise F-statistics, and stepwise mutation model (SMM) index. Gene diversity was calculated at each locus as

$$2n(1-\sum_{u} p_u^2)/(2n-1-f),$$

where p_u is the frequency of the u^{th} allele, n is the sample size, and f is the inbreeding coefficient estimated from genotype frequencies (Weir 1996). SMM index was defined as the maximal proportion of alleles that follow a stepwise mutation pattern (Matsuoka *et al.* 2002a). Analysis of molecular variation (AMOVA) was performed (Excoffier 1992).

To evaluate the probability that each of 260 inbreds would have a unique genotype (fingerprint) for a given number of SSRs, 10,000 random samples of 260 lines were drawn from the empirical distribution of allele frequencies based on the observed data for our 260 inbreds. For these random samples, the probability that all 260 simulated lines had a unique genotype was directly estimated for different numbers of loci. To compare

the relationship of pedigree distance and phylogenetic distance, we used a Mantel test (Mantel 1967) by setting the permutation number to 100,000. Pedigree distances were calculated as 1 - Malécot coefficient of coancestry (Malécot 1948) using pedigree information from a variety of sources (see Supplemental Table S1).

Analysis of genetic structure

Lines were subdivided into genetic clusters using a model-based approach with the software package STRUCTURE (Pritchard *et al.* 1999). Given a value for the number of subpopulations (clusters), this method assigns lines from the entire sample to clusters in a way that Hardy-Weinberg disequilibrium and linkage disequilibrium (LD) were maximally explained. We excluded seven popcorn lines and five sweet corn lines in this analysis (see Results). At least 6 runs of STRUCTURE were done by setting the number of populations (K) from 1 to 10. For each run, burn-in time and replication number were both set to 500,000. The run with the maximum likelihood was used to assign lines to clusters. Lines with membership probabilities ≥ 0.80 were assigned to clusters; lines with membership probabilities < 0.80 for all groups were assigned to a "mixed" group. The three largest clusters were then further subdivided by the same method.

To construct a phylogenetic tree, we used the log-transformed proportion-of-shared-alleles distance that is free of the step-wise assumption, enjoys low variance, and is widely used with multilocus SSR data (Matsuoka *et al.* 2002b). We used the Fitch-Margoliash least squares algorithm implemented in the computer program Phylip to construct phylogenetic trees (Felsenstein 1993). The tree was rooted using five samples

of the maize wild relative, teosinte (*Zea mays* ssp. *parviglumis*) as the outgroup (Matsuoka *et al.* 2002b).

Analysis of allelic richness

We wanted to compare the allelic richness in maize inbreds to that in open-pollinated landrace (exotic) accessions to estimate the extent to which our set of 260 inbreds captures the diversity present in maize overall. For comparison, we used a previously published dataset for exotic maize of 193 samples that represent the entire maize germplasm pool (Matsuoka *et al.* 2002b). To make the comparison of allelic richness of inbreds to exotics, we need to adjust for the inbreeding coefficient since inbreds are mostly homozygous while exotics have a high degree of heterozygosity. We also need to adjust for sample size since our sample has 260 inbreds but only 193 exotics. We used two approaches. First, we compared sets of randomly chosen lines from the inbred and exotic datasets with the same sample sizes. The inbred and exotic genotypes were first broken into alleles to simulate the selfing process. Then, the allele number was counted for randomly drawn samples of size three to 193 in steps of five. Second, we used a parametric simulation to simulate the creation of 260 inbreds from the exotic lines. The inbreeding coefficient (f) for inbreds was estimated to be 0.965. For each locus, we sampled two alleles with replacement to generate a diploid genotype. If the two alleles are the same, then the simulated inbred is made homozygous. If the two alleles are different, then the simulated inbred is made heterozygous with probability 1-f and made *a/a* with probability $f/2$ and *b/b* with probability $f/2$. This procedure was repeated to create 10,000 independent samples of 260 inbreds from which the mean number of alleles

and other summary statistics were calculated. The summary statistics for these simulated data were compared with the actual inbred data.

Estimating the historical sources for inbreds

In order to estimate the historical sources for each inbred group, we used SSR data for 104 representative accessions from four likely historical germplasm pools: Southern dent, Northern flint, Tropical highland maize, and Tropical lowland maize (Supplemental Table S3; Matsuoka *et al.* 2002b). We calculated the likelihood of the allelic constitution of an inbred group (e.g. NSS) or specific inbred line given different proportions of ancestry from the four historical germplasm pools. Assuming that the loci are independent, the likelihood is

$$Lik(p \mid n_{lj}, f_{klj}) \propto \prod_{k=1}^{4} \prod_{j=1}^{a_l} (\sum_{k=1}^{4} p_k . f_{klj})^{n_{lj}},\ 0 < p_k < 1, \sum_{k=1}^{4} p_k = 1$$

Where a_l is number of alleles at the *lth* locus, f_{klj} is the frequency of the *jth* allele at the *lth* locus in the *kth* population as estimated from the 109 representative exotics/landrace lines. n_{lj} is the count of the *jth* allele at the *lth* locus for the inbreds group (or line). p_k is the probability that the allele originated from the *kth* population. This function was maximized by Sequential Quadratic Programming. Several starting points were chosen to check the global convergence. Standard deviation and confidence interval were inferred from the likelihood surface using established methods (Edward 1972).

Defining core sets of inbreds

We developed a new algorithm for building core sets of germplasm by maximizing allelic richness using simulated annealing (Kirkpatrick, Gelatt and Vecchi 1983). Given the complete set of lines (*L*), the algorithm works by first randomly selecting a subset of lines (*l*). Each line has a weight (*w*) based on the number of private alleles in that line. Next, between 1 and the minimum(*l*, *L*-l) additional lines are chosen from the remainder of the complete set (unselected lines) based on their weights and swapped with the same number of the initially selected *l* lines also chosen on the basis of their weights. The number of alleles (*n*) is then evaluated and the swap is accepted if it increases *n* but accepted only with some probability (*P*) if *n* is the same or less. The probability of acceptance is dependent on level of decrease in allelic richness and on the iteration number such that *P* is larger in earlier iterations. Swapping is continued for a predefined number of iteration. Since *P* gradually decreases with iterations (time), the method simulates an annealing process. Under this approach, lines with more private alleles have a larger probability to be included in the core set. Our algorithm can also incorporate a weight for the agronomic quality of the inbred and can allow some inbreds to be designated as "conserved" such that they are automatically included in the core set. The details of the algorithm will be given in a separate paper (See Chapter 3).

Linkage Disequilibria

The matrix of *P*-values for the pair-wise estimates of LD among all 94 SSR loci was evaluated in PowerMarker by the permutation version of Fisher's exact test (Liu and Muse 2003). The numbers of locus pairs with LD *P*-values less than threshold values of 0.05, 0.01, 0.001 were counted for the observed data. This analysis was performed on

each of the clusters of inbreds defined by the program STRUCTURE as well as the entire set of inbreds. Because the *P*-value of the exact test is affected by sample size and the clusters varied widely in size, we evaluated the effect of sample size on the proportion of significant LD *P*-values by drawing random samples from the entire set of 260 lines of a size equal to the actual number of lines in the cluster. These random samples were drawn without replacement, the exact test was performed on each, and the mean proportion of significant LD *P*-values for 100 replicates was used to compare with the actual results for each cluster.

RESULTS

SSR diversity

We surveyed 260 diverse maize inbred lines using 94 SSR loci. The inbreds can be roughly grouped as including 82 tropical lines, 35 temperate Stiff Stalk lines, 131 temperate non-Stiff Stalk lines, seven popcorn lines, and five sweet corn lines. The pedigrees for each line are too extensive to be reported here, but are available online (Supplemental Table S1). Among the lines, we detected a total of 2039 alleles or an average of 21.7 alleles per locus (Table 2.1). There is a large number of private alleles (556 or 27%) that are found in only one of the 260 inbred lines. Most alleles are in low frequency (Figure 2.1).

The number of alleles is not equivalent among loci. Loci with dinucleotide repeat motifs have considerably more alleles (average=23.9) than loci with repeat motifs of three nucleotides or larger (average=9.9; Table 2.1). This difference is also seen for genetic diversity, with dinucleotide SSRs (average=0.839) having a higher genetic diversity than longer-repeat SSRs (average=0.707). The mean genetic diversity of all SSRs is 0.818.

SSRs are often presumed to follow a stepwise mutation process due to changes in the number of repeats. However, because size differences among alleles are estimated based on the combined molecular weights of the SSR plus its flanking sequences, indels in the flanking sequences can contribute to allelic variation as well (Matsuoka *et al.* 2002a). These indels can cause a violation of the expectation that allele sizes differ strictly by multiples of the repeat motif. We calculated a stepwise mutation index, or the maximal

proportion of alleles at a locus that are simple multiples of the repeat motif length. For all 94 loci, the average stepwise mutation model index was 0.832 with dinucleotide SSRs (0.853) showing a higher index than other repeat loci (0.720).

The large number of alleles per locus and the common occurrence of private alleles suggest that a relatively small number of SSRs would be sufficient to uniquely fingerprint maize inbreds. For the 260 inbred lines that we sampled, the following six loci form a unique profile: *bnlg244*, *bnlg2238*, *bnlg619*, *bnlg1191*, *bnlg1046* and *dupssr28*. Assuming the allele distribution of our inbred data is representative of all maize inbreds, the probability of sampling 260 independent lines without generating the same genotype for any two lines will be >0.99 by randomly selecting 10 loci. This number is nine if one only uses dinucleotide SSR and 12 if one uses longer-repeat SSR. Thus, very few SSRs are necessary if one wishes to uniquely fingerprint maize inbreds.

Genetic structure of inbred lines

We wished to assess the degree of relatedness among lines and to identify clusters of genetically similar lines. To do this, we used a model-based approach with the program STRUCTURE to subdivide the lines into clusters (Pritchard *et al.* 2000). Five sweet corn lines and seven popcorn lines were assigned to two pre-defined groups (Sweet and Popcorn) and were excluded in the STRUCTURE analysis. This was done because a pilot analysis showed that incorporating these 12 lines in the analysis interfered with the ability of STRUCTURE to converge on a robust solution. K (number of populations) $= 3$ was found to converge well and showed a comparable or higher likelihoods than

$K = 4$ to 10 among runs of program. We used the run with highest log likelihood at $K = 3$ for the observed data to produce model-based groups (Table 2.2).

The model-based groups are largely consistent with known pedigrees of the lines (MMG and JSCS, pers. obs.). The largest group has 94 lines most of which are regarded by breeders as temperate non-stiff stalk lines (NSS). The next group has 58 lines most of which are either tropical or semitropical lines (TS). The smallest group has 33 lines, all of which are temperate stiff stalk lines (SS). The remaining 63 lines have less than 80% membership in any one group and were assigned to a mixed group. Supplemental Table S3 shows the proportional membership for these mixed lines in the three groups. Most mixed lines are either NSS-TS or NSS-SS mixtures. There are only four lines (Tzi16, Tzi25, Hi27, CML92) that present high membership of TS and SS.

STRUCTURE analysis was repeated to break the three main clusters into subclusters (Table 2.2). The SS group split into four subgroups, the TS group into five, and the NSS group seven subgroups. Supplemental Table S5 shows the proportional membership of the lines in the subgroups for the group to which they belong.

A Fitch-Margoliash "phylogenetic" tree was constructed to further assess the genetic structure of maize inbreds (Figure 2.4). The tree shows good agreement with the pedigree information and STRUCTURE analysis (see Discussion). A version of the tree with the names of the inbreds is available on line (Supplemental Figure S1).

Genetic diversity within inbred groups

Gene diversity and mean numbers of alleles for the 94 SSRs were calculated for each group of inbreds (Table 2.1). The TS group is the most diverse with 13.49 alleles per locus and gene diversity of 0.81. NSS has less diversity than TS, which was revealed by the decreased allele number (12.84) and gene diversity (0.78). SS was found to be less diverse than NSS and TS. Our samples of Sweet and popcorn includes only a few lines, and thus, the small allele numbers in these groups were expected. In all groups, dinucleotide loci have a much larger allele number than longer-repeat loci. Gene diversity also shows a similar trend.

Maize inbreds show a high number of line-specific (556 or 27%) or group-specific (765 or 38%) alleles (Table 2.1). Far more line- and group-specific alleles are found in the TS group (204 and 305) than in the NSS group (121 and 173) despite a much smaller sample size for TS, indicating far greater diversity in tropical than in temperate inbreds.

An AMOVA revealed that most (90.16%) of the genetic variation resides within groups and only a small percentage resides between groups (8.32%) or within lines (1.51%). Overall F_{st} among groups is 0.086 (95% CI 0.080-0.092) with Fst for each locus ranging from 0.02 to 0.17. Pair-wise comparisons show a low level of differentiation between TS and NSS groups (F_{st} =0.06), but more substantial differentiation between SS and the other groups (Table 2.3). Popcorn is also highly differentiated from all the other groups. A similar pattern of differentiation among groups is seen using Nei's minimum distance.

Allelic richness of maize inbreds

Comparison of diversity in inbreds to that in open pollinated landraces shows that the latter possess much greater diversity. For the exotics, the number of allele (2697 or 28.7 alleles per locus) and overall gene diversity (0.84) are higher than for the inbreds (2039 or 21.7; 0.82). To compare allelic richness in inbreds vs. landraces for equal sample sizes, we randomly selected equal numbers of samples from both germplasm pools (see Materials and Methods). This analysis reveals the greater allelic richness in exotics when the samples are equivalent (Figure 2.2). When the sample size is small (<20), the inbreds capture about 88% as many alleles as the exotics. When the sample size is large (>100), inbreds capture an about 78% as many alleles as the exotics.

We also compared allelic richness in inbreds vs. exotics using a parametric simulation. Simulated samples of 260 inbred lines drawn from the landrace gene pool had an average gene diversity of 0.837 (standard error = 0.0015), which is very close to value for the landrace sample (0.840). The minimal value of gene diversity in the simulations is 0.832, which is still higher than our actual inbred sample (0.820). The mean number of allele numbers obtained by the simulations is 2292 and standard error is about 15. The total number for the inbreds sample (2039) is not in the 99% confidence interval [2239, 2334], indicating that if one randomly created a set of 260 inbreds from the exotic gene pool it would contain substantially more allelic diversity than our actual set of 260 inbreds.

Relationship of inbreds to exotic lines

To understand the relationship between the inbreds and exotics, we estimated the proportion of each inbred group's gene pool that was derived from four different segments of the exotics gene pool (Northern Flint, Southern Dent, Tropical Lowland and Tropical Highland). TS has its origin mostly from tropical lowland (66%) and tropical highland (18%) (Table 2.4). NSS and SS show very similar origins, being composed of roughly equal proportions of Northern Flint, Southern Dent and Tropical Lowland. Popcorn has a high proportion of Northern Flint germplasm (40%) with the rest of its genome coming mostly from Tropical Lowland (26%) and Southern Dents (23%). Sweet corn has the largest contribution from Northern Flint germplasm (72%). Overall, Tropical Highland maize has made a modest contribution to our set of inbreds than have the other three historical sources. Variances for these estimates are usually small (SD < 1%). Estimates of historical sources for individual inbreds are included in Supplemental Table S4.

Comparison of SSR and pedigree relationships

A Mantel test shows a highly significant ($p < 10^{-6}$) correlation between pedigree and SSR distance, although the correlation coefficient is relatively small ($r = 0.57$). A plot of pedigree by SSR distances shows a general strong relationships but with many outliers (Figure 2.3).

Core sets of inbreds

We defined core sets of inbreds that capture the maximum number of alleles for a given sample size (Table 2.5). In selecting these sets, we constrained the selection to include six lines (A632, B37, B73, C103, Mo17, Oh43) of high agronomic importance. We also eliminated 8 lines (A654, B2, CM37, CMV3, CO109, I205, Q6199, R109B) because of poor agronomic quality under our field conditions. Additional core set of different sizes can be found in Supplemental Tables S5 and S6. Our study shows 10 lines can capture 28% of all the 2039 SSR alleles in the 260 lines, 20 lines capture 46% of the alleles, 30 lines capture 58%, and 50 lines capture 73%. In order to cover all the possible 2026 alleles, 193 lines were needed. The core sets generally include a large proportion of TS lines as expected since TS have the greatest allelic richness.

Linkage disequilibria

We assessed extent of LD among SSRs for our sample of inbreds. LD was significant at a 0.01 level between 66% of the SSR marker pairs when all lines were included in the analysis (Table 2.6). The proportion of significant pairwise LD tests was less within each model-based group. Reduced power to detect LD with fewer lines could contribute a part of this reduction. However, when we evaluated the percentage of significant pairwise tests in sets of randomly chosen inbreds of the same size as a given group, we observed that sample size alone fails to explain all the reductions (Table 2.6). This suggests that either linkage or population structure within the NSS, TS and SS groups contributes to LD. In particular, SS shows a much larger observed LD value, which may be a consequence of the fact that the SS group actually consists of four well-defined subgroups.

DISCUSSION

SSR Diversity

Previous studies have shown that maize contains abundant SSRs (Senior *et al.* 1993a, 1993b, 1998) and that these SSRs are highly polymorphic even among small samples of maize inbreds (Chin *et al.* 1996; Taramino and Tingey 1996). These pioneering studies were conducted using relatively small numbers of inbreds (9 to 94) and loci (6 to 70). We have extended these earlier analyses by using both a large number of SSRs (94) and a much larger number of inbreds (260) that encompass a much greater portion of the maize gene pool. Our analyses uncovered abundant allelic variation with an average of 21.7 alleles per locus over 94 loci. This value greatly exceed the previously reported values of 5.21 (Senior *et al.* 1998), 6.6 (Taramino and Tingey (1996), 4.9 (Lu and Bernardo 2001), and 6.9 (Matsuoka *et al.* 2002) alleles per locus. The larger number of alleles observed in the present study can be attributed to the larger number of inbreds surveyed, the more diverse selection of inbreds (tropical, subtropical and temperate), and the inclusion of more dinucleotide repeat SSRs, which tend to be more polymorphic than SSRs with longer repeat motifs (Vigouroux *et al.* 2002).

We have also observed higher values of gene diversity than seen in previous analyses of SSR variation in maize inbreds. Gene diversity for our sample of SSRs and inbreds was 0.82 as compared to values of 0.59 (Senior *et al.* 1998), 0.76 (Taramino and Tingey (1996), 0.59 (Smith *et al.* 1997) and 0.62 (Matsuoka *et al.* 2002). Since estimates of gene diversity are not affected by differences in sample size, the higher value that we observed is likely a function of our use of a greater portion of dinucleotide repeat SSRs and of our

more diverse set of inbred. If one considers only SSRs with trinucleotide or longer repeat motifs, then gene diversity in our sample (0.71) falls within the range of these previous reports.

We also showed that most maize SSRs generally fit a stepwise mutation model with 83% of the alleles fitting multiples of the length of the repeat motif of their respective loci. The 17% of alleles that deviate from the stepwise pattern likely represent cases where there have been indels in the regions flanking microsatellite repeat (Matsuoka *et al.* 2002a). The failure of these SSRs to fit exactly a stepwise model cautions against the use of models that assume a stepwise mutation process. In particular, estimates of genetic distance such as $(\delta\mu)^2$ (Goldstein 1995) or measures of population subdivision based on the stepwise mutation model (Slatkin 1995) would be inappropriate to apply to our data.

Genetic structure

Maize inbreds have a complex history, having been derived from multiple open-pollinated varieties and crosses among the inbreds themselves (Gerdes *et al.* 1993). This history makes it difficult to place maize inbreds into realistic groups that reflect their degree of genetic affinities. Pedigree information provides a useful guide, however selection and genetic drift during inbreeding can cause considerable discrepancies between pedigree and genetic constitution. Moreover, pedigree information for some inbreds is either incomplete, inaccurate or conflicting.

We used the model-based approach of Pritchard *et al.* (2000) to define natural groups of maize inbreds. In performing this analysis, we discovered that the inclusion of small numbers of sweet (five) and popcorn (seven) lines in the analysis prevented the convergence to a robust solution. Apparently, these two groups were represented by too few lines to form distinct clusters, while at the same time they are too divergent from the other lines to fit into the clusters for those lines. Only when the sweet and popcorn lines were excluded did STRUCTURE converged on a robust solution with three clusters representing the temperate stiff stalks (SS), other temperate non-stiff stalks (NSS) lines, and tropical-subtropical (TS) lines. Thus, along with the predefined sweet and popcorn lines, we classify maize inbreds into five groups. Sixty-three lines did not fit into one of these five groups since they consist of a mixture of two or more of the primary groups. A comparison of genetic distances among the five groups indicates that SS are the most divergent (Table 2.3), a result consistent with the observation that the SS lines typically provide a strong heterotic response in crosses with other maize inbreds (Hallauer *et al.* 1988).

Inbreds in each of the three model-based groups were analyzed again using STRUCTURE to identify subclusters of related lines (Table 2.2). The SS group split into four tight subgroups of lines derived from B14, B73, B37, and N28 (see Anonymous 1999). The TS group split into five distinct subgroups with a clear relationship to the origin of these lines. For example, lines in the TZI subgroup are fully tropical and many are from the IITA's streakbreeding-resistance program. The Suwan subgroup consisted mainly of tropical lines that were derived from the Suwan-1 composite population,

principally of Caribbean origin (Sriwatanapongse 1993), plus B96 from Maíz Amargo of Argentina. The CML-late subgroup is comprised of tropical lines tracing back to CIMMYT's late-maturing Tuxpeño composite populations. The CML-early subgroup contained lines derived from CIMMYT's early-maturing (in the tropics) Tuxpeño related materials and other intermediate-maturity sources. The NC subgroup consists of lines derived from Latin American tropical hybrids. Lines in the subgroup CML-P were largely derived from the La Posta Population 43 developed at CIMMYT from 16 lines of Tuxpeño origin (CIMMYT 1987).

The NSS group is organized into seven subgroups that reflect known heterotic groups (Anonymous 1999). The lines in subgroup Hy:T8:Wf9 all trace to these three lines that were important in the era of double-cross hybrids. Lines in subgroup M14:Oh43 all trace to M14, which was an important inbred in the 1940-50's, and Oh43, which is still among our most important breeding sources. Several lines (A556, MS1334, ND246 and W401) with no known relationship to one another were grouped loosely within the CO109:Mo17 subgroup. Mo17 and CO109 represent important, but independent, breeding sources. Subgroup C103 consists mostly of Lancaster germplasm (Hallauer et al. 1988) with an additional contribution from B57. Lines within subgroup Ga:SC (Georgia and South Carolina) are mostly southern US germplasm with the notable exception of line 4226. Subgroup NSS-X is a heterogeneous mixture of mostly older lines. The K64W subgroup contains a set of related white lines.

A Fitch-Margoliash tree based on the SSR data shows generally good agreement with the pedigree information and STRUCTURE analysis (Figure. 2.4, Supplemental Figure S1). There is a general separation of the TS, NSS and SS lines. Mixed lines are usually located between clusters of TS/NSS/SS lines. Within the SS lines, the four subgroups defined by the STRUCTURE analysis are perfectly matched with four clades. For the TS group, the tree has three clades that correspond to subgroups NC, Suwan, and CML-late. For the NSS lines, the tree has three clades that largely correspond to subgroups Hy:T8:Wf9, M14:Oh43 and K64W. All of the sweet corns fall in the same clade, as did all of the popcorns. The European (F2, F7, EP1) lines and one Canadian (CO255) line are closely grouped together, and this clade is neighbor to the sweet corn clade as expected since all these lines were derived from the Northern Flint landrace of the northern US and adjacent Canada (Galinat 1971; Doebley *et al.* 1985). NSS-X is also contained within a single large clade, despite the fact that these lines have heterogeneous pedigrees.

Genetic diversity among inbred groups

The amount of genetic diversity within each of the model-based groups is not equivalent. Rather, gene diversity is highest in tropical inbreds (TS) followed by NSS, sweet corn, SS and popcorn in that order. The greater diversity of the TS lines is again shown by the fact that TS lines contain more alleles than NSS (1268 vs. 1207) despite the fact that the sample size for TS was much smaller (58 vs. 94). TS lines also possess by far the greatest number of group-specific alleles (305). These data argue strongly that TS inbreds represent an important source of diversity for broadening the genetic base for maize breeding (Goodman 1985; Goodman and Carson 2000).

Five hundred and fifty-six of the 2039 alleles (27%) occur in only one inbred, and 765 alleles (38%) are restricted to a single model-based group of inbreds. These large proportions of private alleles are probably a function of the high mutation rate for maize SSRs (Vigouroux *et al.* 2002), allowing much new allelic variation to arise within lines after their initial development. This feature of maize SSRs contributes to their considerable discriminatory power, enabling one to fingerprint uniquely our entire set of 260 lines with as few as ten SSRs. This discriminatory power makes SSRs ideal markers for use in varietal identification (Smith *et al.* 1997) and for monitoring gene flow between lines (Dale *et al.* 2002). SSRs can also be used to determine pedigrees in maize inbreds and hybrids but more (e.g. 60 or more SSR loci) are required to trace pedigrees than to provide for unique line identification especially when closely related inbreds are considered (Berry *et al.* 2002).

We also compared diversity in maize inbreds relative to the open-pollinated landraces from which the inbreds were ultimately derived. For this purpose, we used a sample of 193 landrace accessions that represent the entire maize germplasm pool. In particular, we examined the number of alleles captured in our set of 260 inbreds as compared to the number of alleles expected to be captured if these 260 lines represented a random sample of the maize gene pool. The results, whether obtained by a random sampling approach or parametric simulation, revealed a deficit of alleles within the 260 inbreds relative to expectations. For example, a set of 260 inbreds selected at random from the maize gene pool would be expected to capture 2292 alleles based on the parametric simulations while the actual set of 260 lines captures only 2039. This result argues that plant breeders could

capture additional diversity by working with landrace accessions (Goodman 1985). It is likely that the landraces contain numerous agronomically useful alleles not represented in the inbreds and advanced populations with which breeders presently work.

Historical sources of maize inbreds

To better understand the relationship between our set of 260 inbreds and the broader maize germplasm pool from which they were derived, we made maximum likelihood estimates of the portions of four segments of the landrace gene pool (Northern Flint, Southern Dents, Tropical Lowland and Tropical Highland) represented in the five inbred groups. The results are consistent with historical records, pedigree information and geography. The temperate NSS and SS are composed of a near-equal mix of Tropical Lowland, Southern Dent and Northern Flints, although the Northern Flint portion is a bit smaller. Since Southern Dents themselves are thought to have been recently derived from tropical lowland germplasm (Galinat 1985; Doebley *et al.* 1988), the high portion of Tropical Lowland germplasm in NSS and SS lines likely represents a tropical contribution that came via the Southern Dents. The observation that NSS and SS are composed of only 25% Northern Flint is consistent with prior observations (Doebley *et al.* 1988).

In addition to our estimates of historical contributions to the inbred groups, we have estimated the historical sources for each of our individual 260 inbreds (Supplemental Table S4). The only inbred in our sample with a high proportion of tropical highland germplasm (72%) is CML349, which is a tropical highland inbred line. This again points

to the possibility of using tropical highland germplasm to increase diversity within maize inbreds. The top four lines in terms of Northern Flint contribution (IA2132, IL14H, IL101t, P39) are all sweet corn. Some European lines (F2, F7) and one Canadian line (CO255) also have more than 50% Northern Flint origin. Va35, a southern US line, was found to have the largest Southern Dent proportion (63%).

Pedigree vs. genetic distance

Previous studies using molecular markers have generally shown a strong correlation between molecular marker and pedigree-based distance measures (Smith and Smith, 1992; Bernardo *et al.* 1997; Smith *et al.* 1997; Bernardo *et al.* 2000; Bernardo and Kahler 2001). Nonetheless, estimates of relatedness on the basis of pedigree data can differ from those based on molecular marker data (Bernardo *et al.* 2001). Calculations of relatedness based upon pedigree data are dependent upon the assumptions that both parents contribute an equal number of alleles, (i.e. no selection, mutation or genetic drift) and that the pedigree data are accurate. Another assumption is that founder genotypes (genotypes for which no further pedigree information on ancestors is available) are unrelated by pedigree. All of these assumptions can be violated.

We observed a highly significant correlation between pedigree and SSR based distances, although a much weaker correlation (0.57) than seen in some previous studies. For example, Smith *et al.* (1997) reported a correlation of 0.81 between SSR and pedigree distances for maize inbreds. Since their sample of inbreds included many commercial lines with detailed pedigrees, it is not surprising that they observed a stronger correlation

than we did with our more diverse set of lines. Similarly, Bernardo *et al.* (2000) observed a correlation of 0.92 between SSR and pedigree using a small set of public inbreds with well-documented pedigrees. Our study also differs from these two prior studies in using a higher proportion of dinucleotide SSRs, which with their higher mutation rate, could weaken the correlation between SSR and pedigree distance.

Linkage disequilibrium

Overall, 66% of SSR pairs exhibited significant LD. Smaller percentages of SSR pairs showed significant LD within the model-based groups, due in part to reduced statistical power with the smaller sample sizes. Sets of inbreds chosen at random from the full set of 260, but of the same size as one of the model-based groups, showed less LD than the actual model-based groups themselves (Table 2.6). This result suggests that the observed LD is largely due to either population structure (or linkage) within groups as opposed to higher level population structure among the entire set of lines. In particular, we observe a large excess of SSR pairs in LD within SS (29%) as compared to the expected number (4%), suggesting either considerable population structure or linkage effects among SS lines. Curiously, our results are in disagreement with those of Remington *et al.* (2001) who found that higher level structure makes an important contribution to LD and who observed the least evidence for LD within SS lines. Differences in sampling might explain these discrepancies.

Perspective

There is a heightened awareness of the necessity for maintaining genetic diversity for the study of natural variation and for crop improvement. However, when stocks are placed in germplasm banks without an adequate understanding of the amount and distribution of genetic variation within those stocks, potential users of these resources are confronted with the difficulty of choosing a diverse and representative selection from long lists of essentially anonymous accessions. In this paper, we have shown that maize inbreds possess a great depth of allelic diversity. This diversity is not distributed randomly among the lines, but rather diversity is structured into five groups along breeding group (SS vs. NSS) and ecological (temperate vs. tropical) axes. Similarly, the amount of diversity is not equivalent among groups, but rather tropical-subtropical inbreds possess greater diversity than their temperate counterparts. It is also clear that allelic diversity in some portions of the broader maize gene pool are not wellrepresented in available inbreds. In particular, we found that the diversity in tropical highland maize is poorly represented among available inbreds, suggesting that tropical highland germplasm could be tapped to identify new alleles of agronomic importance. Finally, to aid researchers working with maize, we have defined both core sets of maize inbreds and a method for choosing core sets to best represent diversity among a set of inbreds. These results should help maize researchers to make more informed choices of inbreds for research and breeding.

ACKNOWLEDGEMENTS

This work was supported by the U.S. NSF grant DBI-0096033. We thank Bruce Weir for comments on the manuscript.

REFERENCES

Anderson E, and Brown WL, Origin of Corn Belt maize and its genetic significance, pp. 124-148 in *Heterosis - A record of researches directed toward explaining and utilizing the vigor of hybrids*, edited by J. W. Gowen. Iowa State College Press, Ames, IA (1952).

Anonymous, *MBS Genetics Handbook*. Mike Brayton Seeds Inc., Ames, Iowa (1999).

Austin DF, Lee M and Veldboom LR, Genetic mapping in maize with hybrid progeny across testers and generations: plant height and flowering. *Theor Appl Gen* **102**:163-176 (2001).

Bernardo R, Breeding potential of intra- and interheterotic group crosses in maize. *Crop Science* **41**: 68-71 (2001)

Bernardo R, and Kahler K, North American study on essential derivation in maize: inbreds developed without and with selection from F2 populations. *Theor. Appl. Genet* **102**: 986-992 (2001).

Bernardo R., Murigneux A, Maisonneuve JP, Johnsson C, and Karaman Z, RFLP-based estimates of parental contribution to F2- and BC1-derived maize inbreds. *Theor. Appl. Genet*. **94**: 652-656. (1997)

Bernardo R, Romero-Severson J, Ziegle J, Hauser J, Joe L, Hookstra G, and Doerge R, Parental contribution and coefficient of coancestry among maize inbreds: pedigree, RFLP, and SSR data. *Theor Appl Genet* **100**: 552-556 (2000).

Berry DA, Seltzer JD, Xie C, Wright DL and Smith JSC, Assessing probability of ancestry using simple sequence repeat profiles: applications to maize hybrids and inbreds. *Genetics* **161**: 813-824. (2002)

Burr B, Burr FA, Thompson KH, Albertson MC and Stuber CW, Genetic mapping with recombinant inbreds in maize. *Genetics* **118**:519-526 (1988).

Chin EC, Senior ML, Shu H and Smith JS, Maize simple repetitive DNA sequences: abundance and allele variation. *Genome* **39**: 866-873 (1996).

Ching A, Caldwell KS, Jung M, Dolan M, Smith OS, Tingey S, Morgante M and Rafalski A, SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. *BMC Genet* **3**:19:1-14 (2002).

CIMMYT, CIMMYT report on maize improvement 1982-83, pp. 1-78. Centro Internacional de Mejoramiento de Maíz y Trigo, Chapingo, Mexico (1987).

Crosbie TM, Mock JJ and Pearce R, Inheritance of photosynthesis in a diallel among eight maize inbred lines from Iowa Stiff Stalk Synthetic. *Euphytica* **27**:657-664 (1978).

Dale PJ, Clarke B and Fontes EMG, Potential for the environmental impact of transgenic crops. *Nature Biotech*. **20**: 567-574 (2002).

Doebley J, Wendel JF, Smith JSC, Stuber CW and Goodman MM, The origin of cornbelt maize: the isozyme evidence. *Econ. Bot.* **42**: 120-131 (1988).

Doebley J, Goodman MM and Stuber CW, Exceptional genetic divergence of the Northern Flint corns. *Amer. J. Bot.* **72**: 64-69 (1986).

Edwards AWF, Likelihood: an account of the statistical concept of likelihood and its application to scientific inference. Cambridge University Press, Cambridge (1972).

Edwards MD, Stuber CW and Wendel JF, Molecular-marker-facilitated investigations of quantitative-trait loci in maize. I. Numbers, genomic distribution and types of gene action. *Genetics* **116**: 113-125 (1987).

Excoffier L, Smouse P and Quattro J, Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. *Genetics* **131**: 479-491 (1992).

Felsenstein J, PHYLIP - Phylogeny Inference Package version 3.5c. Department of Genetics, University of Washington, Seattle (1993).

Fowler JE and Freeling M, Genetic analysis of mutations that alter cell fates in maize leaves: dominant Liguleless mutations. *Developmental Genet*. **18**: 198-222 (1996).

Galinat WC, The evolution of sweet corn. University of Massachusetts-Amherst College of Agriculture Agricultural Experiment Station Research Bulletin 591: 1-20 (1971).

Galinat WC, Domestication and diffusion of maize, pp. 245-282 in Prehistoric Food Production in North America, edited by R. I. Ford. University of Michigan, Ann Arbor, MI (1985).

Gerdes JT, Behr CF, Coors JG and Tracy WF, Compilation of North American Maize Breeding Germplasm. *Crop Sci. Soc. Am.*, Madison, Wisconsin. Pages: 1-202 (1993).

Goldstein DB, Ruiz Linares A, Cavalli-Sforza LL and Feldman MW, Genetic absolute dating based on microsatellites and the origin of modern humans. *Proc. Natl. Acad. Sci*. USA **92**: 6723-6727 (1995).

Goodman MM, Exotic maize germplasm: Status, prospects, and remedies. *Iowa State J. Res.* **59**: 497-529 (1985).

Goodman MM and Carson ML, Reality vs. myth: corn breeding, exotics and genetic engineering. *Proc. Annu. Corn Sorghum Res. Conf.* **55**: 149-172 (2000).

Goodman MM, Broadening the genetic diversity in breeding by use of exotic germplasm, pp. 139-148 in Genetics and Exploitation of Heterosis in Crops, edited by J. Liu et al. – page 30 G. Coors and S. Pandey. Crop Science Society of America, Madison, WI (1999).

Hallauer AR, Russell WA and Lamkey K, Corn breeding, pp. 463-564 in Corn and Corn improvement, edited by G. F. Sprague and J. W. Dudley. Crop Science Society of America, Madison, WI (1988).

Henry A and Damerval C, High rates of polymorphism and recombination at the *Opaque-2* locus in cultivated maize. *Mol Gen Genet* **256**:147-157 (1997).

Kirkpatrick S, Gelatt CD and Vecchi MP, Optimization by simulated annealing. *Science* **220**: 671-680 (1983).

Labate JKL, Mitchell S, Kresovich St, Sullivan H, Smith JSC, Molecular and historical aspects of corn belt dent diversity. *Crop Sci* **43**: 80-91 (2003).

Liu K, Powermarker - A powerful software for marker data analysis. North Carolina State University Bioinformatics Research Center, Raleigh, North Carolina (www.powermarker.net) (2003).

Lu H and Bernardo R, Molecular marker diversity among current and historical maize inbreds. *Theor Appl Genet* **103**: 613-617 (2001).

Malécot G, Les mathématiques de l'hérédité. Masson & Cie, Paris (1948).

Mantel N, The detection of disease clustering and a generalized regression approach. *Cancer Res.* **27**: 209-220 (1967).

Matsuoka Y, Mitchell SE, Kresovich S, Goodman MM, Doebley J, Microsatellites in Zea - variability, patterns of mutations, and use for evolutionary studies. *Theor Appl Genet* **104**: 436-450 (2002a).

Matsuoka Y, Vigouroux Y, Goodman MM, Sanchez GJ, Buckler E et al., A single domestication for maize shown by multilocus microsatellite genotyping. *Proc. Natl. Acad. Sci.* USA **99**: 6080-6084 (2002b).

Poethig RS, Heterochronic mutations affecting shoot development in maize. *Genetics* **119**: 959-973 (1988).

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

Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR et al., Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc. Natl. Acad. Sci.* USA **98**: 11479-11484 (2001).

Romero-Severson J, Smith JSC, Zeigle J, Hauser J, Joe L & Hookstra G. Pedigree analysis and haplotype sharing within diverse groups of Zea mays L. inbreds. *Theor. Appl. Genet.*, **103**: 567-574 (2001).

Senior ML, and Heun M, Mapping maize microsatellites and polymerase chain reaction confirmation of the targeted repeats using a CT primer. *Genome* **36**: 884-889 (1993).

Senior ML, Chin ECL, Lee M, Smith JSC and Stuber CW, Simple sequence repeat markers developed from maize sequences found in the GENBANK database: map construction. *Crop Sci* **36**: 1676-1683 (1996).

Senior ML, Murphy JP, Goodman MM, Stuber CW, Utility of SSRs for determining genetic similarities and relationships in maize using an agarose gel system. *Crop Sci* **38**: 1088-1098 (1998).

Slatkin M, A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139**: 457-462 (1995).

Smith JSC, Chin ECL, Shu H, Smith OS, Wall SJ, Senior ML, Mitchell SE, Kresovich S, Ziegle J, An evaluation of the utility of SSR loci as molecular markers in maize (Zea mays L): comparisons with data from RFLPS and pedigree. *Theor Appl Genet* **95**: 163-173 (1997).

Smith OS, and Smith JSC, Measurement of genetic diversity among maize hybrids – a comparison of isozymic, RFLP, pedigree, and heterosis data. *Maydica* **37**: 53-60 (1992).

Sriwatanapongse S, Jinahyon S and Vasal S, Suwan-1: maize from Thailand to the world, pp. 1-16. Centro Internacional de Mejoramiento de Maíz y Trigo, Chapingo Mexico (1993).

Taramino G and Tingey S, Simple sequence repeats for germplasm analysis and mapping in maize. *Genome* **39**: 277-287 (1996).

Tenaillon MI, Sawkins MC, Long AD, Gaut RL, Doebley JF et al., Patterns of DNA sequence polymorphism along chromosome 1 of maize (Zea mays ssp. mays L.). *Proc. Natl. Acad. Sci.* USA **98**: 9161-9166 (2001).

Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D et al., Dwarf8 polymorphisms associate with variation in flowering time. *Nature Genet*. **28**: 286-289 (2001).

Troyer AF, Chapter 14 - Temperate Corn. In Specialty Corn, A. Hallauer (ed.), CRC press (2001).

Vigouroux Y, Jaqueth JS, Matsuoka Y, Smith OS, Beavis WD et al., Rate and pattern of mutation at microsatellite loci in maize. *Mol. Biol. Evol.* **19**: 1251-1260 (2002).

Weir BS, Genetic Data Analysis II. Sunderland, MA: Sinauer (1996).

Table 2.1: Summary statistics for all inbreds and each subgroup

Statistics [a]	Overall	TS	Sweet	NSS	Popcorn	SS	Mixed
Sample size	260	58	5	94	7	33	63
Alleles	2039	1268	272	1207	277	535	1321
Alleles per locus	21.7	13.49	2.89	12.84	2.95	5.69	14.05
Type I SSR alleles/locus	23.9	14.71	2.91	13.91	2.97	5.99	15.22
Type II SSR alleles/locus	9.9	7.07	2.80	7.20	2.80	4.13	7.93
Gene diversity	0.82	0.81	0.64	0.78	0.54	0.59	0.82
Type I SSR gene diversity	0.84	0.83	0.64	0.80	0.56	0.61	0.84
Type II SSR gene diversity	0.71	0.68	0.65	0.68	0.45	0.51	0.72
Group specific alleles	765	305	26	173	16	43	202
Group specific alleles/line	2.94	5.26	5.20	1.84	2.29	1.30	3.21
Group specific alleles (%)		24.05	9.56	14.33	5.78	8.04	15.29
Line specific alleles	556	204	18	121	11	36	166
Line specific alleles (%)		16.09	6.62	10.02	3.97	6.73	12.57

[a] Type I markers are dinucleotide SSRs and Type II markers are SSRs with longer-repeat motifs.

Table 2.2: List of the 260 lines by their model-based groupings

Group[a]	Subgroup[b]	Lines
NSS	Hy:T8:Wf9	CI21E, H49, Hy, Mo1W, Pa875, Pa880, T8, Va17, Va14, Va22, Va35, Va102, W64A, Wf9
	M14:Oh43	A619, Gn2, H95, M14, Oh40B, Oh43, Oh43E, PA762, Va26, Va85
	CO109:Mo17	A556, A682, CI.187-2, CO109, CO220, K187, Mo17, MS1334, ND246, W401
	C103	B57, C103, C123, DE2, L317, L1546, NC258, NC262
	Ga:SC	4226, F44, Ga209, GT112, SC357, SC213R, SC213
	NSS-X	38-11, A239, A659, AR4, CM7, CM37, R168, Mo44, MS71, NC260, PA884P, R4, R177, W22
	K64W	33-16, CI.31A, CI.64, CI.66, CI.7, E2558W, Ky21, K55, K64, M162W
	NSS-mixed	A554, A654, B2, B52, B70, B77, B97, B103, CO106, CO125, F6, Fe2, H99, Mt42, N6, Os420, Pa91, R109B, SD44, T234, W153R
SS	B14A	A214N, A632, A634, A635, A665, B14A, B64, B68, CM105, CM174, H91
	B37	B37, B76, H84, NC250
	N28	N28, N28Ht
	B73	A679, A680, B73, B84, B104, B109, NC328, NC372, R229
	SS-mixed	A641, De811, H100, N192, N196, NC294, NC368
TS	TZI	A6, CML52, CML238, CML287, NC358, Q6199, Tzi8, Tzi9, Tzi10, Tzi18
	Suwan	B96, CML69, CML228, CML349, Ki3, Ki9, Ki11, Ki14, Ki44, Ki2007
	CML-late	CML5, CML9, CML61, CML103, CML220, CML254, CML258, CML261, CML264, CML314, Tx601
	CML-early	CML14, CML247, CML311, CML321, CML322, CML331, CML332
	NC	NC296, NC298, NC304, NC336, NC338, NC348, NC350, NC352, NC354
	CML-P	CML10, CML11, CML45, CML277, CML281, CML333, CML341
	TS-mixed	CML38, CML108, NC300, NC356
Sweet		Ia2132, Il14H, Il101t, Il677a, P39
Popcorn		HP301, I29, IDS28, IDS69, SA24, Sg18, Sg1533
Mixed		A188, A272, A441-5, A656, B79, B94, B105, B164, C49A, CML77, CML91, CML92, CML218, CML323, CML328, CMV3, CO159, CO255, D940Y, DE3, EP1, F2, F2834T, F7, Hi27, I137TN, I205, IDT, Ki43, Ky226, Ky228, L578, Le23, Le773, M37W, Mo18W, Mo24W, Mp339, MS153, N7A, NC264, NC320, NC360, NC362, NC364, NC366, NC370, Oh7B, Oh603, SC55, SD40, SD46, T232, TEA, Tx303, Tzi11, Tzi16, Tzi25, U267Y, Va99, W117, W117Ht, W182B

The 260 lines in our study are listed with the grouping and subgrouping from STRUCTURE analysis. Lines in the mixed group show less than 80% membership for

any group. Seven popcorn lines and five sweet corn lines were assigned into predefined popcorn and sweet corn groups. Within TS, SS, and NSS group, an additional subclustering organizes each group into several distinct subgroups and one mixed subgroup by using the same scheme.

[a] The groups are SS - stiff stalk lines, NSS - non stiff stalk lines, TS - tropical/semitropical lines, sweet corn, popcorn, and mixed lines (see text).

[b] The subgroups are named after a defining inbred line(s), principal source (e.g. NC for North Carolina), maturity (early vs. late), or mixed for lines that showed less than 80% membership for any subgroup.

Table 2.3: Genetic distances between maize inbred groups

Group	TS	Sweet	NSS	Popcorn	SS
TS	-	0.58	0.29	0.52	0.47
Sweet	0.15	-	0.47	0.62	0.61
NSS	0.06	0.12	-	0.46	0.32
Popcorn	0.15	0.29	0.15	-	0.57
SS	0.18	0.28	0.14	0.31	-

Upper triangle is Nei's-minimum distance and the lower triangle is pairwise F_{ST} .

Table 2.4: Historical sources for each maize inbred group

Group	Tropical Lowland (mean ± s.e.)	Southern Dents (mean ± s.e.)	Tropical Highland (mean ± s.e.)	Northern Flints (mean ± s.e.)
NSS	0.31±0.01	0.37±0.01	0.05±0.01	0.27±0.01
Popcorn	0.26±0.03	0.23±0.02	0.11±0.03	0.40±0.03
SS	0.32±0.01	0.38±0.01	0.08±0.01	0.23±0.01
Sweet	0.14±0.02	0.06±0.02	0.08±0.02	0.72±0.03
TS	0.66±0.01	0.11±0.01	0.18±0.01	0.04±0.01

The estimates and their standard errors of the historical sources are summarized for each group. These are MLEs. Variance was estimated from the observed Hessian matrix. Because of MLE's asymptotical normality, the 95% confidence interval can be constructed approximately from mean ± 1.96*s.e.

Table 2.5: List of core sets of inbred lines

Sample size	Alleles obtained	Line list
10	579	A632, B37, B73, C103, Mo17, Oh43, CML5, Tzi18, CML91, CML52
20	943	A632, B37, B73, C103, Mo17, Oh43,, CML14, CML277, CML52, Tzi8, M37W, CML281, CML228, Oh7B, Il14H, CML322, CML91, B96, Tx601, Mo18W
30	1179	A632, B37, B73, C103, Mo17, Oh43, B96, Tzi8, CML277, CML228, Ky21, Mo18W, Oh7B, CML5, CML322, CML220, A441-5, ML61, Tx303, CML14, CML91, CML311, CO159, CML281, Il101t, Tx601, CO255, A272, M37W, CML77
50	1481	A632, B37, B73, C103, Mo17, Oh43, CML77, CML261, IDS28, CML277, B96, CML14, CML322, CML91, Mo18W, CML220, CML281, I137TN, Ky21, CML228, CML5, Tzi8, A272, A441-5, W401, Oh7B, CML349, CML69, Hi27, F2, CML61, P39, Tzi9, CML247, CI.7, CML254, NC364, CML328, Il14H, CO159, CML321, OS420, Va85, NC304, Tx303, CML311, NC348, M37W, B57, K55

The first 6 lines (A632, B37, B73, C103, Mo17, Oh43) were conserved because of their agronomic importance. A654, B2, CM37, CMV3, CO109, I205, Q6199, R109B were excluded because of poor agronomic performance in our fields in Raleigh and Florida.

Table 2.6: Percentage of SSR locus pairs in LD at a $p = 0.01$ level

Population	No. of lines	Observed % in LD	Expected % in LD[a]
Overall	260	66.05%	
NSS	94	19.29%	18.91%
TS	58	14.48%	9.13%
SS	33	28.92%	4.32%

[a] Based on average percentage of all locus pairs showing LD in a random samples containing the same number of lines.

FIGURE LEGENDS

Figure 2.1: Histogram of allele frequency. There are 2039 alleles in total. There is also a large number of the alleles (265 or 13%) that are at very low frequency (<0.01) although present in more than one inbred. One thousand-forty alleles (56%) are present at frequencies between 0.01 and 0.25, 72 alleles (3.5%) have frequencies between 0.25 and 0.50, 5 alleles (0.2%) have frequencies between 0.50 and 0.75, and only one allele (0.05%) has frequency above 0.75.

Figure 2.2: Plots of allele number obtained against sample size. For a given sample size, 1000 replicates were sampled from the inbreds dataset or exotics dataset without replacement and the genotypes were randomly broken into alleles. Then the mean number was calculated to give the plot from sample size 3 to 193. Log trendlines (not shown) fit the plots very well.

Figure 2.3: Plot of the proportion of shared SSR alleles distance between inbred lines by the pedigree distance between inbreds. Pedigree distance is defined as 1-Malecot Coefficient of Coancestry (Malecot 1948).

Figure 2.4: Fitch-Margoliash tree for the 260 inbred lines using the log transformed proportion of shared allele distance. The tree was rooted using five teosinte (*Z. mays* ssp. *parviglumis*) samples as outgroups.

Figure 2.1

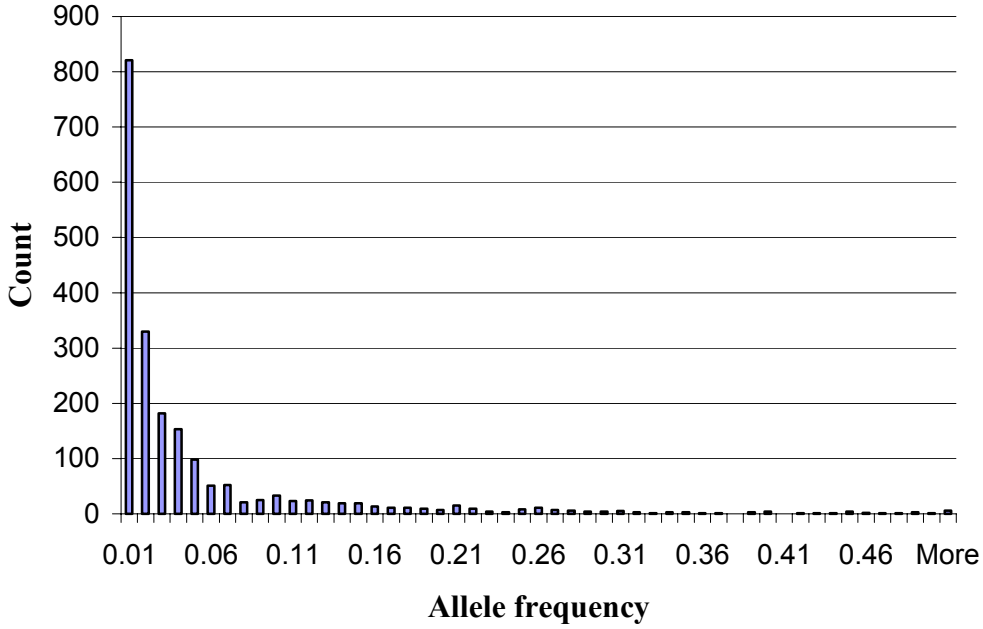

900
800
700
600
500
400
300
200
100
0
Count
0.01
0.06
0.11
0.16
0.21
0.26
0.31
0.36
0.41
0.46
More
Allele frequency

Figure 2.2

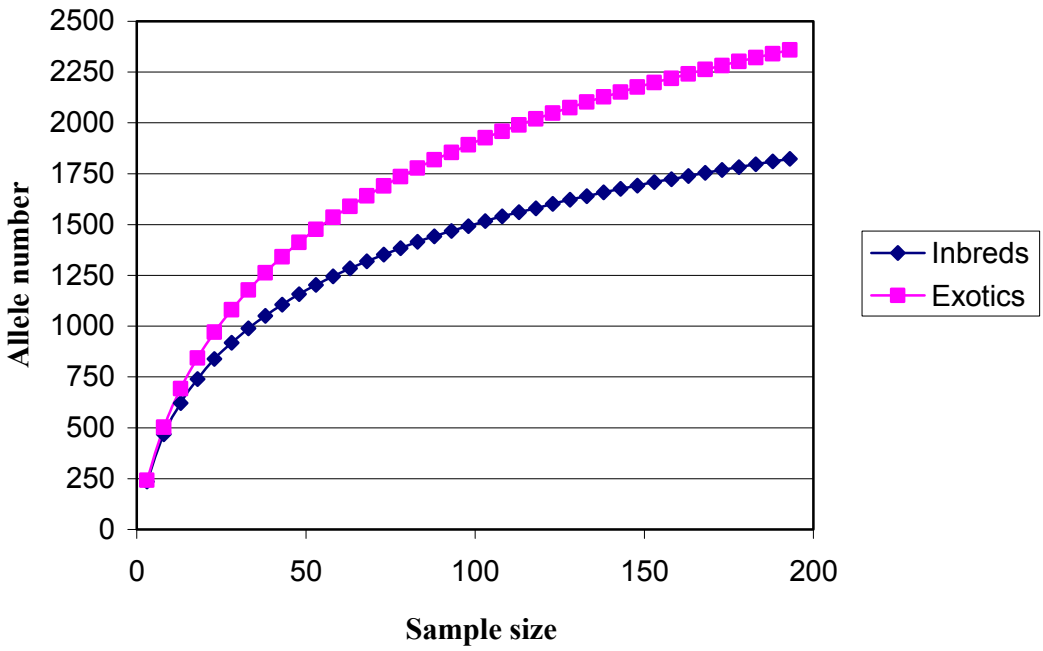

2500
2250
2000
1750
1500
1250
1000
750
500
250
0
Allele number
0
50
100
150
200
Sample size
Inbreds
Exotics

Figure 2.3

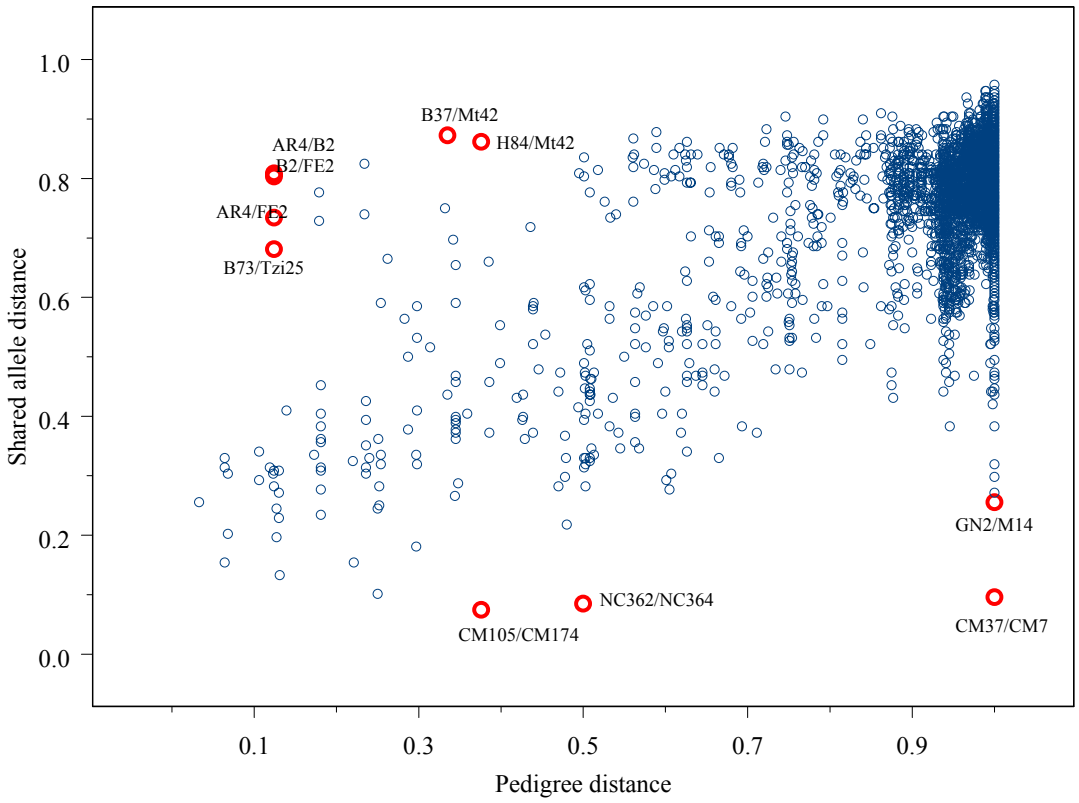
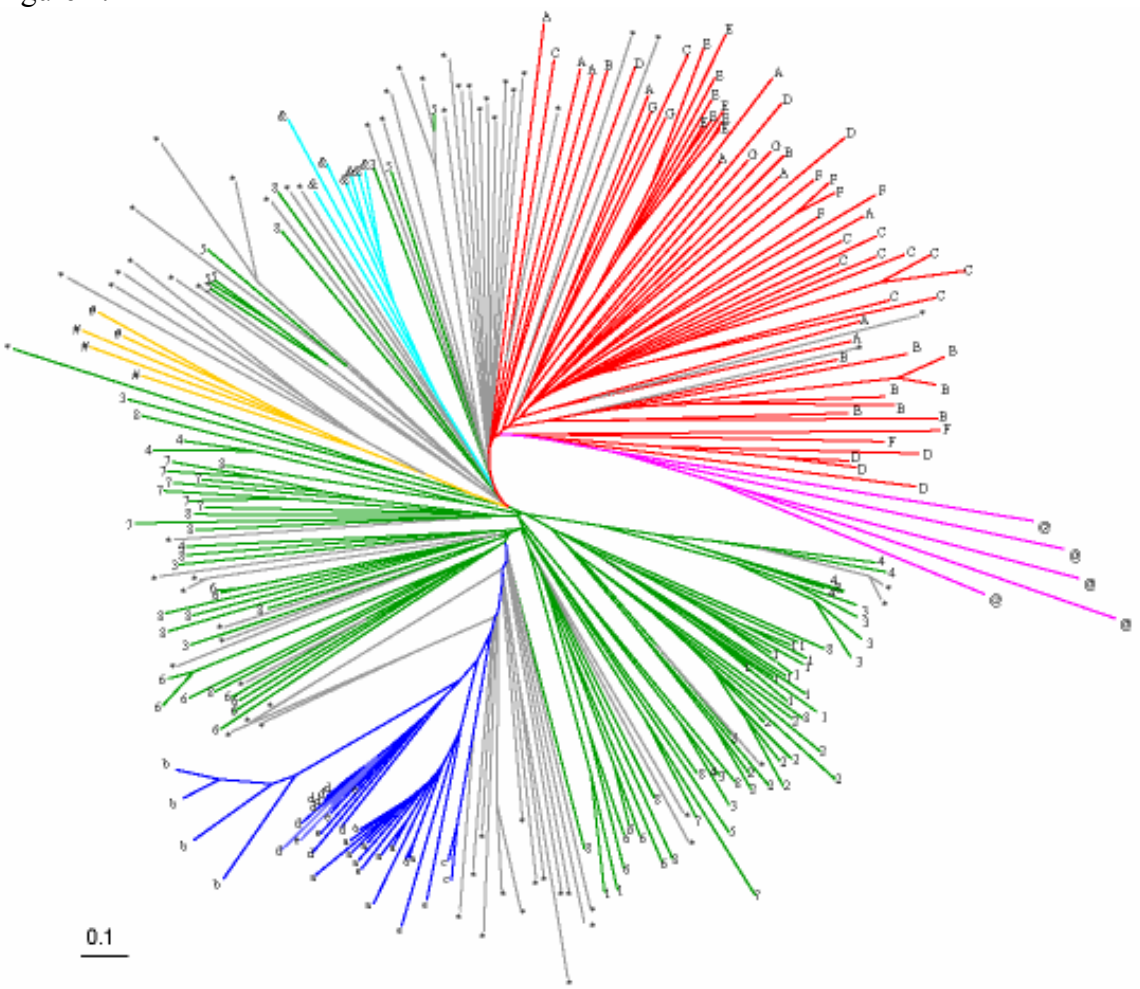

Shared allele distance
1.0
0.8
0.6
0.4
0.2
0.0
0.1
0.3
0.5
0.7
0.9
Pedigree distance
B37/Mt42
H84/Mt42
AR4/B2
B2/FE2
AR4/FE2
B73/Tzi25
GN2/M14
NC362/NC364
CM105/CM174
CM37/CM7

Figure 2.4



NSS group
1 Hy:T8:Wf9
2 M14:Oh43
3 CO109:Mo17
4 C103
5 Ga:SC
6 NSS-X
7 K64W
8 NSS-mixed

SS group
a B14A
b B37
c N28
d B73
e SS-mixed

TS group
A TZI
B Suwan
C CML-late
D CML-early
E NC
F CML-P
G TS-mixed

Mixed group *
Popcorn &
Sweet #
Outgroup @

Chapter 3

Choosing Core Sets of Lines from a Large Germplasm Pool

Kejun Liu[1], Xiang Yu[1], and Spencer Muse[1,2]

1. Bioinformatics Research Center, North Carolina State University

   Raleigh, NC 27695

2. Correspondence:

   Dr Spencer V Muse

   501 Partners II, Bioinformatics Research Center

   North Carolina State University

   Raleigh, NC 27695

   Phone: (919) 515-1948

   Fax: (919) 515-7315

   Email: muse@stat.ncsu.edu

ABSTRACT

In this paper we describe a simulated annealing (SA) algorithm for choosing a core set of lines from a large germplasm collection. Any measure of core set quality can be used in the algorithm. The most important feature of our algorithm is that general constraints can be incorporated in the selection. The algorithm can be used to find the minimal set of lines with maximal diversity, or given a sample size find the optimal core set of all possible sets. Our algorithm is computationally efficient and adjustable. With weak convergence conditions the algorithm finds nearly optimal local maxima very quickly. Under strong convergence conditions global maxima are almost guaranteed. Our algorithm illustrates the versatility of simulated annealing on handling combinatorial algorithms. The simulated annealing algorithm is applied to analyze the maize inbreds germplasm of 102 inbreeding lines. Under the constraint of including at least one line originating from each of the 5 major groups, the optimal 50 lines include 90% of all the alleles, whereas on average 70 randomly picked lines will be required to obtain the same allelic richness. The algorithm can easily be extended to the general question of choosing the optimal K-subset from a large sample space with constraints.

INTRODUCTION

There are a variety of settings where it is necessary to select a subset of lines, populations, or individuals to represent some larger set of germplasm. For instance, conservation geneticists may want to identify a "core set" of populations that maximize the total amount of genetic variation, subject to an upper limit on the number of lines they can maintain (Frankel 1984; Frankel and Brown 1984; Schoen and Brown 1983). Plant or animal breeders may wish to select a subset of available breeding lines or populations for breeding stock maintenance purposes (Brown 1989a; Gouesnard *et al.* 2000). Recognizing these objectives, Frankel and Brown (1984) defined a core set of lines as a subset of a larger germplasm collection that maximizes the possible genetic diversity with a minimum of repetitiveness. In this paper, we describe algorithms for selecting such a core set in a way that maximizes diversity for a given user-defined number of lines.

In the absence of detailed genetic information about lines, core sets have typically been selected by first stratifying the entire germplasm collection on the basis of phylogeny, geography, or phenotypic characters, followed by selection of core set lines in numbers proportional to the groups they represent (Brown 1989b). In recent years, genetic diversity in an enormous number of species has been at least partially characterized thanks to the rapid decrease in monetary and time costs associated with genotyping technologies including allozymes, restriction fragment length polymorphism (RFLPs), microsatellites or simple sequence repeats (SSRs), and single nucleotide polymorphisms (SNPs). Schoen and Brown (1993) proposed the so-called M (maximization)-strategy for identifying core sets with maximum genetic diversity. The M-strategy simply enumerates

all the possible core sets of a given size, and then singles out the sets with the maximum total observed number of alleles at the surveyed loci. The key assumption of the M strategy is that observed allelic richness at the genotyped marker loci is correlated to the allelic richness at unobserved loci. Monte Carlo simulations of germplasm collections have shown that the M-strategy performs well under a variety of genetic models (Bataillon *et al.* 1996). The M-strategy is especially useful for inbreeding species because of the genome-wide correlation of genetic variation resulting from a decrease in the spatial decay of linkage disequilibrium (Schoen *et al.* 1993).

The basic M-strategy approach outlined above is certainly effective when the number of possible core sets under consideration is small. However, an exhaustive search of all possible core sets is practically infeasible when selecting a moderate to large core set of size k from a large germplasm pool of N lines. To remedy this practical problem, Bataillon *et al.* (1996) proposed an iterative search to give an approximation solution. First, a randomly chosen subset of k lines is selected as the initial core set. Next, each of the lines is deleted in turn to form a series of collections of size $k-1$, and the collection having maximal allelic richness (defined simply as the total number of observed alleles) is retained. Each of the remaining $N-k$ lines are then added, forming a series of sets of size k, and the set with maximal richness is retained. This process of removing one line, then replacing it with the line that maximizes allelic richness, is repeated until either the procedure converges (the removed line is replaced with itself) or a predetermined maximum number of iterations is reached.

While this iterative search appears to be effective and provides an approximate solution to the core set selection problem, it does not guarantee an optimal solution. In fact, the nature of the algorithm's swapping is likely to result in a local maximum rather than a global maximum. Furthermore, it is difficult to incorporate constraints on the composition of the core set collection under the Bataillon *et al.* strategy. In practice, it is often the case that a variety of constraints are placed on the final core set. As a simple example, suppose that the N lines in the germplasm collection have been stratified into several geographically distinct groups. It might be desirable to select a core set of k lines subject to the constraint that at least 3 lines from each geographic group are in the final set. It is not obvious how to incorporate such constraints in the Bataillon *et al.* algorithm. Below, we present an algorithm for finding constrained core sets that have globally maximum levels of a chosen genetic diversity measure.

The task of choosing the core sets is a combinatorial optimization problem. Simulated annealing has gained wide acceptance as a general algorithmic approach to solving hard combinatorial optimization problems in a variety of settings (Kirkpatrick 1983, Brooks and Morgan 1995). The major difficulty for implementing a simulated annealing approach for the core set problem is ensuring that the user-specified constraints are satisfied. While the work presented below targets the specific problem of how to choose a core set of lines from a large germplasm collection, the methods and results apply to the more general question of how to choose the (perhaps constrained) subset of k items from a larger set of N items so that a criterion function is optimized.

METHODS

Given an existing germplasm collection of N lines with genetic polymorphism data, X, available, there are two prerequisites for choosing a subset, λ_k, of k lines to form a small, informative core set. First, an objective function, $f(\lambda_k;X)$, must be chosen for measuring the information content of each subset. For the purpose illustrating the algorithms, we will use the total allele number of the subset as the objective function f (see Discussion for other useful measures). Having decided upon f, the goal is to find a subset of k lines, λ_k^*, that furnishes a maximum value of f from the set of all possible subsets of k lines. Exhaustive enumeration of all possible $\binom{N}{k}$ k-subsets is infeasible for most applications. For instance, with 50 lines, there are 1.2×10^{22} possible core sets of size 12.

Unconstrained Optimization

For the unconstrained case, a general simulated annealing algorithm was applied to choose a core set of lines. While the application of simulated annealing is straightforward in this setting, we outline it briefly in Appendix A to provide a framework for describing the solution to the constrained case.

Constrained Optimization

In practice, a variety of constraints must be incorporated in the core set selection procedure. Important lines should be conserved, and representatives of each of several pre-specified groups (identified by such factors as geography, phylogeny, or phenotype) may be required in the core set. We use a hierarchical approach to model the constraints.

Suppose the germplasm collection is organized into several non-overlapping subsets, each falling into one of the four possible types: unconstrained, constrained, conserved, excluded. In our chosen core set we require all n_C lines from the conserved subset S_C, and none of the n_E lines from the excluded subset S_E. Lines from the unconstrained subset S_U may be included or excluded as needed. Each constrained subset, S_i, $i = 1,2,...,m$, has upper and lower bounds U_i and L_i, and the number of the n_i lines in S_i appearing in the final core set must fall between the bounds.

The challenging aspect of incorporating the constraints in a simulated annealing algorithm is producing a computationally efficient mechanism for exploring only valid core sets that satisfy all the constraints. In Appendix B we describe an algorithm for achieving this goal.

Multiple Selection Criteria

One might want to carry out the selection of core sets based not simply on genetic diversity information. Instead, there might be multiple criteria, both genetic and non-genetic, that could contribute to the choice of lines. Examples include various genetic measures, population size, phenotypic information, ease of line maintenance, economic costs, or growing season. A naive approach for incorporating multiple criteria is to use the diversity information and the simulated annealing algorithms to report the top, say, twenty core sets. Secondary criteria can be computed for each of these subsets, with an "optimal" one chosen by the user. Clearly this method is *ad hoc* and guarantees no sort of optimality. A more general and rigorous approach involves computing a (linear)

combination of criteria values for each line, much like the notion of a selection index in breeding settings. The value of this composite function then replaces the allelic richness value as our optimality criterion. Because different criteria have different ranges of possible values and follow different distributions, criteria should be normalized before the function is computed. If prior knowledge or theory provides appropriate normalization functions, they can be used. In the absence of such information, we take an empirical approach to the normalization process. A random sample of *k*-subsets is selected, and the values for each criterion are computed. The sample mean and standard deviation is then used to estimate the mean and standard deviation of the distribution for each criterion. The composite objective function is defined as $\sum_j w_j \frac{f_j(\lambda_k;X)-\hat{\mu}_j}{\hat{\sigma}_j}$, where w_j is the (user-provided) weight for *j*th criterion, $f_j(\lambda_k;X)$ is the observed value of the chosen *k*-subset for the *j*th criterion, and $\hat{\mu}_j$ and $\hat{\sigma}_j$ are the estimated means and standard deviations of the *j*th criteria from the random sampling process described above.

Program

Within the PowerMarker package (Liu and Muse 2003) we implemented a module called CoreSet to automate the selection process. The module supports a batch script for the specification of user-defined parameters and constraints that is useful for automated or high-throughput settings. For simple use of the CoreSet module, the program also provides a "wizard" to generate the script from the graphical user interface. The PowerMarker package is available at http://www.powermarker.net.

RESULTS

Performance evaluations

We used a simple test data set to compare the performance of new and existing algorithms for selecting core sets. The test data set contains the 29 non stiff stalk inbred lines of maize from Matsuoka *et al.* (2002), each of which has been typed for 94 microsatellite markers (a larger set of 102 lines is used later in the Results section). We used total allele number as our objective function, and investigated the ability of methods to identify core sets varying in size from 2 to 10. The small number (29) of lines was chosen to allow for exhaustive evaluation of all possible core sets. In Table 3.1 we present a comparison of three different core set selection algorithms: (i) exhaustive search, (ii) the iterative search of Bataillon *et al.* (1996), and (iii) the simulated annealing algorithm of this paper. Simulated annealing was performed under a weak convergence condition (see Appendix A). The number of evaluations for each annealing schedule, R, was set to 100, and the cooling coefficient ρ was set to 0.8. Since the results of the simulated annealing and iterative searches rely on a random component, the values vary over replicate analyses of the same data. Thus, we report the means and standard deviations of the function values from 1000 replicate analyses of the 29 lines. To indicate the amount of time consumed by each method, we report for each method M, the total number of objective function evaluations required for a single replicate.

The average value of the objective function is the most relevant piece of information for comparing the different methods. Note that in every instance, the average value found by the simulated annealing algorithm is very close to the true value achieved by the

exhaustive search. In the worst case ($k = 2$) the core set found by simulated annealing differed from the true maximum by an average of only 1.7, demonstrating that the algorithm on average recovers a core set within less than 1% of the true maximum value. For the most computationally challenging case ($k = 10$), the simulated annealing algorithm result is, on average, only 0.3% below the true value. In comparison, the average result of the Bataillon et al iterative algorithm falls more than 9% below the true maximum when $k = 10$. In no case did the Bataillon *et al.* approach outperform simulated annealing. Also important is the replicate-to-replicate variation of the core sets, described by the standard deviations from the 1000 replicate analyses. Note that simulated annealing has roughly half the variation of the Bataillon *et al.* method.

In the first column of Table 3.1 we see the rapid growth in evaluation number of the exhaustive search that necessitates the iterative and simulated annealing algorithms. Note that moving from core sets of size size 2 to core sets of size 10, the total evaluation number of the exhaustive search increases approximately 50,000-fold. In contrast, the evaluation number of simulated annealing increased only around 50%, while the Bataillon *et al*. algorithm showed an increase of 145%.

In Figure 3.1 we examine the effectiveness of the Bataillon *et al.* and simulated annealing algorithms in finding the global maximum, again using the 29 maize inbred lines. 1000 replicate analyses were performed for each core set size, $k = 2..10$. Simulated annealing was evaluated with the swapping number for a single schedule (R) of both 100 and 500. The Bataillon *et al.* method was allowed to iterate until it converged for 10 successive

steps. The vertical axis in Figure 1 shows the percentage of replicate analyses in which the different methods achieved the true global maximum (found by exhaustive search). For example, when searching for a core set of 7 lines, simulated annealing with strong convergence conditions ($R = 500$) evaluations reached the global maximum 100% of the time; using weak convergence conditions ($R = 100$) the global maximum was found about 80% of the time. The Bataillon *et al.* algorithm found the global maximum in this case less than 20% of the time. In only one case did simulated annealing with strong convergence condition ($R = 500$) evaluations fail to recover the true maximum at a rate greater than 95% (For the case of $k = 6$, where multiple local maxima differ from the global maximum by only one allele). Simulated annealing outperformed the Bataillon *et al.* method in each case, with comparable computational expense (The number of core set evaluations for iterative search is larger than that of the simulated annealing with $R = 100$ but less than that of the simulated annealing with $R = 500$ **)**

Finally, we investigated the performance of the constrained simulated annealing algorithm. We selected line Mo17 to be conserved, and SC213R to be excluded. At least 1 but no more than 2 lines from (38-11, A554, A619, B103, B97, C103, CI187-2, CM7, F44) were to be in the core set, as were no more than 3 of (H95, H99, I29, K55, Ky21, M162-W) (see Table 3.3). Table 3.2 shows that the algorithm is extremely effective in recovering the optimal core sets under this setting.

Improving performance with weighted sampling

In the algorithms described in Appendices A and B, we do not specifically describe the random sampling procedure we use to select lines from the available set. In the simple case, we sample lines uniformly. Efficiency can be improved by selecting lines from the available set according to a non-uniform distribution. The use of a weighted sampling scheme will not change the behavior of global convergence, but can improve the convergence speed significantly. The weights associated with each line are dependent on the specific criterion function, but should be assigned in way that more promising core sets will get evaluated with higher probability. When the objective is to maximize the total allele number of the core set, we have found that weights for each line based on the private allele number prove to be effective (the private allele number for a line is simply the number of alleles present only in that line). Private alleles will always increase the total allele number, so lines with many private alleles are preferred. Similar weighting schemes can be developed for other criteria.

Selecting a core set of maize inbreds

We used our simulated annealing algorithm to select core sets from the 102 maize inbred lines of Matsuoka *et al.* (2002), each of which was genotyped at 100 microsatellite markers. The germplasm collection was partitioned into six groups based on a model-based clustering approach (Pritchard *et al.* 2000). Table 3.3 lists the names and groupings of the 102 lines. We used the total allele number as the objective function to maximize. The parameters for the simulated annealing algorithm were set to $R = 500$ and $\rho = 0.9$. The simulated annealing algorithm was then used under the constraint that at least one

line from each of the five major groups was included in the final core set. The Mixed group was not constrained. Core sets ranging from size $k = 5$ to $k = 100$ were found, in steps of 5. We replicated the analysis 10 times for each core set size, and the results are shown in Figure 2 and in Table 3.4. In Figure 3.2 we plot a curve showing the percentage of the total allele number in the germplasm collection of 102 lines captured in core sets ranging in size from 5 to 100. We see that a core set of 50 lines includes almost 90% of all alleles found at the 100 microsatellite loci in these 102 lines. In Table 3.4 we present the lines comprising core sets of size 5, 10, 20, 30, and 50. The core set selection algorithm was replicated independently 10 times for each of these cases, and the reported sets were found in at least 9 of the 10 replicates, providing some confidence that these are indeed the globally optimal core sets. It is important to note that the replication was carried out simply to provide evidence for the reliability of the identified core sets; such replication would be optional in a practical setting.

DISCUSSION

The simulated annealing algorithm developed in this paper has been shown to be a more effective and efficient means for selecting core sets of lines than existing published methods. This result is perhaps not surprising, given the versatility of simulated annealing in providing effective solutions to hard combinatorial optimization problems (Kirkpatrick *et al.* 1983; Goradia and Lange 1988; Lukashin *et al.* 1992). An important practical decision in the use of this algorithm, or, indeed, in the use of any algorithm for selecting core sets based on genetic data, is the choice of a measure describing core set quality. We used total allele number for illustrative purposes in this paper, but other measures may be more appropriate in some settings. For example, one might want to include allele frequency information if the inclusion of rare alleles is not of particular value. In this case, the use of allelic diversity as a criterion considering both allele number and frequency would be a superior choice. An advantage of the simulated annealing method is the ease of incorporating non-genetic data into the selection criteria. Care must be taken when combining these data types, and normalization procedures are imperative. Our work suggests that simple adjustments based on forming "z-scores" from the normal distribution are usually sufficient. However, it is advisable to carry out empirical experiments investigating randomly chosen core sets to check for substantial deviations from normality.

The computational problem addressed in this work is an example of the *minimum test set* problem, which has been shown to be NP-complete (Garey 1979). A variety of other problems in genetics, as well as in other settings, are examples of this problem, and our

algorithm can be expected to provide good practical solutions. As an example, scientists are interested in finding the minimum number of of representative SNPs within a genomic block that uniquely distinguish all haplotypes. Published studies rely on exhaustive enumeration of all SNP combinations (Clayton *et al.* 2001; Zhang *et al.* 2002). While this is an ideal solution when the haplotype block has a small number of SNPs, it becomes impractical as the number of SNP sites increases. The application of our simulated annealing algorithm using haplotype diversity as a criterion function provides an effective solution to this problem.

ACKNOWLEDGEMENTS

We thank John Doebley, Major Goodman, Edward Buckler for helpful discussions.

APPENDIX A: Unconstrained simulated annealing

Algorithm 1: Simulated annealing for choosing an optimal unconstrained k-subset

We assume the reader has a basic knowledge of simulated annealing, as described in Kirkpatrick (1983) or Brooks and Morgan (1995).

Step 1: Starting a new annealing schedule with initial temperature T_0, we randomly pick an initial *k*-subset, λ_k with function value $(\lambda_k ; X)$, where X denotes the data.

Step 2: Propose a new *k*-subset, λ_{new}, by swapping a randomly chosen line from λ_k with a line not in the core set, the swap being accepted with probability

$$\Pr(\text{replace } \lambda_k \text{ with } \lambda_{new}) = \begin{cases} 1, & D \geq 0 \\ e^{D/T}, & D < 0 \end{cases},$$

where $D = f(\lambda_{new}; X) - f(\lambda_k; X)$.

Step 3: Repeat Step 2 *R* times. If any core set changes occurred during those *R* steps, we say that a move was made. If the function value of the core set increased, we call the move "upwards"; if the function value decreased, we call the move "downwards". (Note that it is possible for changes to occur and for the move to be neither upwards nor downwards.)

Step 4: Update the annealing temperature to ρT, where T is the annealing temperature of the previous step and ρ, $0 < \rho < 1$, is the cooling coefficient.

Step 5: Repeat steps 2-4 until one of the following three conditions are satisfied: (1) No successful moves were made in a single annealing schedule. (2) A user-defined maximum number of annealing schedules was reached. (3) All the moves in a single annealing schedule are neither "downwards" nor "upwards".

Condition (1) is the regular stopping rule for general simulated annealing. As no successful move was made, the system is considered to have converged, and the algorithm stops. Condition (2) was added to guarantee an answer in finite time. In some special cases, two or more global maxima with the same values of the objective function can make the system circulate in an infinite loop, so condition (3) was included to monitor this behavior. When all the moves are neither downwards nor upwards, the system is assumed to be in a cycling state, and the algorithm stops. We have not observed this behavior in practice.

The probability of finding the global maximum and the time taken to find it are determined by the three parameters R, ρ, and T_0. A large value of the swapping number, R, for a single annealing schedule increases the probability of finding the global maximum at the expense of run time. Likewise, larger values of ρ provide slower, but more accurate, algorithms. A sufficiently large value of T_0 is required to encourage the algorithm to explore the potential solutions effectively, but if it is too large then too much time is spent before the system cools. It is recommended that several values of these parameters be tried before deciding on a suitable value. For general use, we set $R = 1000, \rho = 0.95, T_0 = 1$.

APPENDIX B: Constrained simulated annealing

In Appendix A we demonstrated that each cycle of the simulated annealing algorithm involved proposing a new core set, λ_{new}, and comparing it with the current core set λ_k. The modifications for the constrained version of the algorithm center on two tasks: selecting an initial random core set that satisfies the constraints; given a valid core set, moving to a valid random "neighbor" core set. The first task is simple, the second is less obvious. We detail algorithms for each below.

Algorithm 2: Choose an initial core set satisfying constraints

We will build a valid set S_{in} by progressively adding lines from the set of available lines, S_{out}. The final collection, S_{in}, will be used as the initial λ_k in the simulated annealing algorithm.

Step 1. Initialization: Let $S_{in} = S_C$; let $S_{out} = S_U + S_1 + S_2 + ... + S_m$.

Step 2. Insure lower bounds are satisfied: For $i = 1, 2, ..., m$, randomly select L_i lines from S_i for inclusion in S_{in}. Denote the selected lines from S_i as $S_{in,i}$ and those unselected as $S_{out,i}$; update the included and available sets:

$$S_{in} = S_{in} + S_{in,1} + S_{in,1} + ... + S_{in,m};$$

$$S_{out} = S_{out} - S_{in,1} - S_{in,2} - ... - S_{in,m}.$$

Step3. Insure upper bounds are satisfied: For each constrained subset S_i randomly choose $n_i - U_i$ lines from $S_{out,i}$ and remove those lines from S_{out}.

Step 4. Complete initial core set: S_{in} now consists of $\text{size}(S_{in}) = \text{size}(S_C + S_{in,1} + ...S_{in,m})$ lines. Remove $k - \text{size}(S_{in})$ lines at random from S_{out}, and add them to S_{in} to complete a valid core set of *k* lines.

The second necessary algorithm is one for choosing a valid new random core set falling within a neighborhood of a valid existing core set.

Algorithm 3: Move from a valid core set, λ_k, to a valid neighbor core set, λ_{new}.

Step 1: Initialization. $S_{in} = \lambda_k$, $S_{out} = S - S_E - S_{in}$, where *S* is the complete set of *N* lines.

Step 2: Select a line to swap out of λ_k: Randomly select a line, s, from λ_k.

Step 3: Determine valid lines for swapping with s and adjust S_{out} accordingly:

i. If $s \in S_i$ and $\text{size}(S_{in,i}) = L_i$, set $S_{out} = S_{out,i}$

ii. For each i, $i = 1,2,...,m$ for which $s \notin S_i$, if $\text{size}(S_{in,i}) = U_i$ then set $S_{out} = S_{out} - S_{out,i}$

Step 4: Complete the new set: Randomly select a line t from S_{out}. Add line t to S_{in}, remove line s from S_{in}

The algorithm for the constrained case follows the same steps as the unconstrained case, with Algorithms 2 and 3 replacing Steps 1 and 2 in Algorithm 1.

REFERENCES

Bataillon TM., David JL, and Schoen DJ, Neutral genetic markers and conservation genetics: simulated germplasm collections. *Genetics*, **144**, 409-417 (1996).

Brooks SP and Morgan BJT, Optimization using simulated annealing, *The Statistician*, **44**, 241-257.

Brown AHD, A case for core collections. In: The use of plant genetic resources (Brown AHD, Frankel OH, Marshall DR, and Williams JT, eds). Cambridge: Cambridge University press; 136-156. (1989a).

Brown AHD, Core collections: a practical approach to genetic resources management: *Genome*, **31**, 818-824 (1989b).

Clayton D., Choosing a set of haplotype tagging SNPs from a larger set of diallelic loci. ftp://ftp-gen.cimr.cam.ac.uk/software (2001).

Frankel OH., Genetic perspectives of germplasm conservation. In: Genetic manipulation: impact on man and society (Arber W, Llimensee K, Peacock WJ, and Starlinger P, eds) Cambridge: Cambridge University Press; 161-170 (1984).

Frankel OH, and Brown AHD, Plant genetic resources today: a critical appraisal. In: Crop genetic resources: conservation & evaluation (Holden JHW and Williams JT, eds). London: George Allen & Unwin; 249-257 (1984).

Garey MR and Johnson DS, In: Computers and Intractability (Freeman, New York), 222 (1979).

Goradia TM. and Lange J, Applications of coding theory to the design of somatic cell hybrid panels, *Math. Biosci*. **91**, 201-219 (1988).

Gouesnard B, Bataillon TM, Decoux G, Rozale C, Schoen DJ, and David JL, MSTRAT: An algorithm for building germplasm core collections by maximizing allelic or phenotypic richness, *The Journal of Heredity*, **92**, 93-94 (2001).

Kirkpatrick S, Gelatt CD, Vecchi MP, Optimization by simulated annealing, *Science*, **220**, 671-680 (1983).

Liu K and Muse SV, PowerMarker: new genetic data analysis software. Version 1.0. Free program distributed by the author over the internet from http://www.powermarker.net

Lukashin AV, Engelbrecht J, and Brunak S, Multiple alignment using simulated annealing: branch point definition in human mRNA slicing, *Nucleic Acids Res.* **20** (1), 2511-2516 (1992).

Matsuoka Y, Vigouroux Y, Goodman MM, Sanchez JG., Buckler E and Doebley J, A single domestication for maize shown by multilocus microsatellite genotyping. *Proc. Natl. Acad. Sci.* USA **99**:6080-6084 (2002).

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

Schoen DJ and Brown AHD, Conservation of allelic richness in wild crop relatives is aided by assessment of genetic markers. *Proc Natl Acad Sci* USA, **22**: 10623-10627 (1993)

Zhang K, Deng M, Chen T, Waterman MS and Sun F, A dynamic programming algorithm for haplotype block partitioning. *Proc Natl Acad Sci* USA, **99**: 7335-7339 (2002)

Table 3.1: Efficiency of different optimization algorithms

Size	Method							
	Exhaustive search		Iterative search			Simulated annealing		
	Eval. #[a]	Value[b]	Eval. #	Mean	S.D.	Eval. #	Mean	S.D.
2	406	205	353.1	202.8	3.0	261	203.3	2.9
3	3,654	278	378.4	275.3	3.1	303.1	276.6	1.7
4	23,751	343	388.2	339.6	3.4	289.7	342.0	1.7
5	118,755	399	418.4	396.6	3.4	330.7	398.4	1.2
6	475,020	450	472.2	446.4	4.5	325.4	448.9	1.5
7	1,560,780	493	537.2	489.2	4.4	333.1	492.5	1.4
8	4,292,145	533	574.5	529.0	4.6	351.8	531.9	1.6
9	10,015,005	570	642.4	565.1	4.9	369.6	569.0	1.9
10	20,030,010	604	865.1	598.4	4.9	349.1	602.6	2.6

[a] Number of evaluations.

[b] Total allele number of the core set across all markers.

Table 3.2: Efficiency of constrained optimization

Size	Method				
	Exhaustive search		Simulated annealing		
	Eval. #[a]	Value[b]	Eval.#	Mean	S.D.
2	9	192	618.2	192.0	0
3	198	273	355.4	272.7	0.9
4	2,025	336	360.7	335.8	0.3
5	12,672	393	329.4	392.6	0.6
6	53,901	443	283.3	442.4	0.9
7	164,538	486	343.9	485.5	0.9
8	372,141	526	318.1	524.9	0.9
9	635,580	561	279.8	560.6	1.1
10	827,739	589	374.1	588.1	1.6

[a] Number of evaluations.

[b] Total allele number of the core set across all markers.

Table 3.3: List of 102 maize inbreds lines

Group	Lines
NSS[a]	38-11, A554, A619, B103, B97, C103, CI187-2, CM7, F44, GT112, H95, H99, I29, K55, Ky21, M162W, Mo17, NC258, NC260, ND246, Oh43, PA91, SC213, SC213R, T8, Va26, W153R, W64A, Wf9
SS[b]	A632, B104, B14A, B37, B68, B73, B84, CM105, CM174, N192, N28Ht, NC250
Sweet	Ia2132, II101, Il14H, Il677a, P39
Popcorn	HP301, IDS28, SA24, Sg18
TS[c]	A6, CML10, CML247, CML254, CML258, CML261, CML277, CML281, CML287, CML333, CML5, CML61, KUI11, KUI2007, KUI21, KUI3, KUI44, NC296, NC298, NC300, NC304, NC338, NC348, NC350, NC352, NC354, Q6199, Tx601, Tzi10, Tzi18, Tzi8
Mixed[d]	A272, A441-5, CML91, CMV3, D940Y, EP1, F2834T, F7, I137TN, I205, KUI43, M37W, Mo24W, MS153, NC320, Oh7B, SC55, T232, U267Y, W117Ht, W182B

The 102 lines in our study are listed with the grouping and subgrouping from STRUCTURE analysis.

[a] Non stiff stalk temperate lines

[b] Stiff stalk temperate lines

[c] Tropical/subtropical lines

[d] Lines showing evidence of multiple origins

Table 3.4: Maize inbreds core sets identified by simulated annealing

Core set size	Percentage of total aleles	CoreSet
5	26.8%	F44, Il677a, KUI21, NC250, SA24
10	43.2%	CML277, CML281, Ia2132, IDS28, KUI21, Ky21, NC250, Oh7B, Q6199, Tzi8
20	67.6%	A272, CML247, CML277, CML281, CML5, CML61, CML91, F44, H99, IDS28, Il677a, KUI21, Ky21, M37W, NC250, Oh7B, P39, Q6199, Tzi18, Tzi8
30	74.2%	38-11, A272, B68, CML247, CML261, CML277, CML281, CML5, CML91, EP1, F44, H99, Il677a, K55, KUI11, KUI21, Ky21, M162W, M37W, Mo24W, NC250, Oh7B, P39, Q6199, SA24, SC213R, SC55, Tzi18, Tzi8, Va26
50	88.9%	38-11, A272, A441-5, B68, B84, CML10, CML247, CML254, CML261, CML277, CML281, CML287, CML333, CML5, CML61, CML91, EP1, F44, F7, H99, HP301, I137TN, Ia2132, Il14H, K55, KUI11, KUI2007, KUI21, KUI43, Ky21, M162W, M37W, Mo24W, NC250, NC258, NC300, NC304, Oh7B, P39, Q6199, SC213R, SC55, T8, Tx601, Tzi10, Tzi18, Tzi8, U267Y, Va26, W153R

Figure legends

Figure 3.1: Comparison of iterative search and simulated annealing. The three vertical bars for each core set size indicate the percents obtaining the global maximum by iterative search, simulated annealing with $R = 100$ and simulated annealing with $R = 500$, respectively. The percents were calculated from 1000 replicates. Global maximum was calculated by exhaustive search.

Figure 3.2: Plots of allele number obtained against core set size. For a given core set size, 10 replicates of simulated annealing were performed on the inbreds dataset with the constraints described in the text. The maximum allele number was calculated to give the plot from sample size 5 to 100.

Figure 3.1

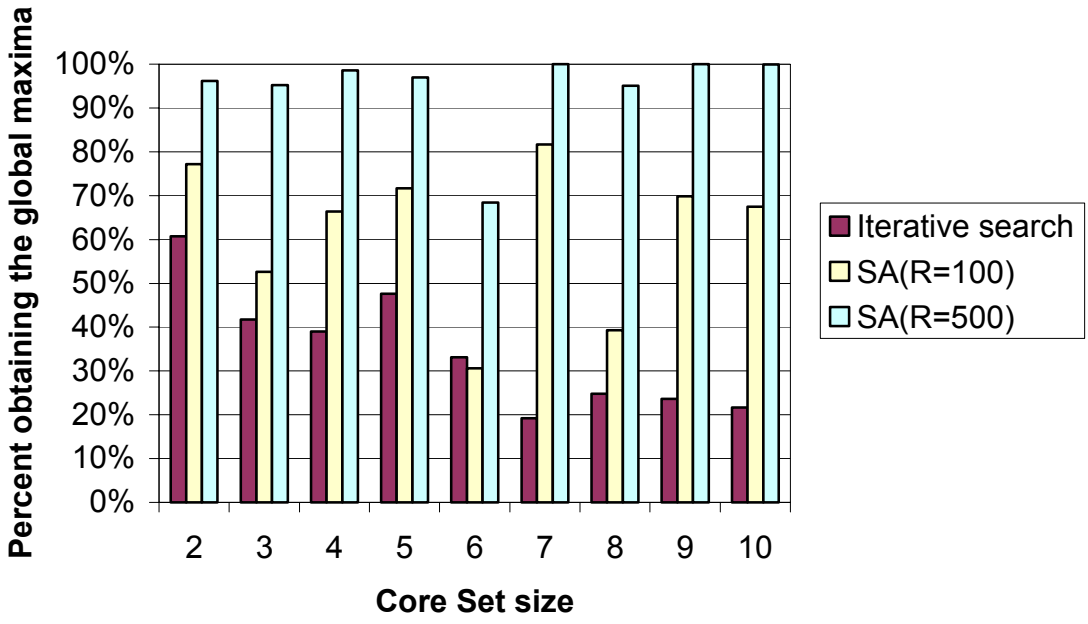
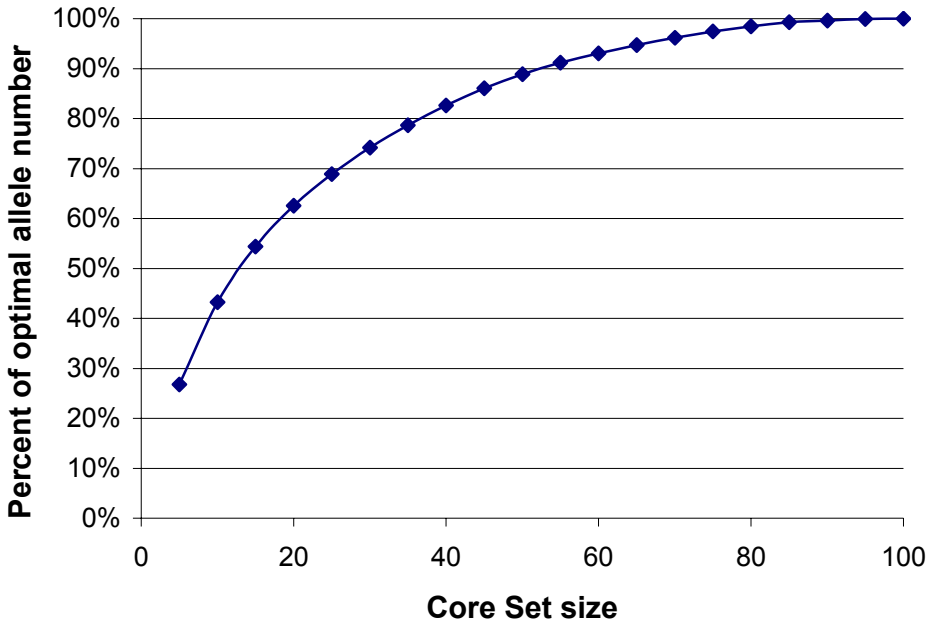

Percent obtaining the global maxima
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
2
3
4
5
6
7
8
9
10
Core Set size
Iterative search
SA(R=100)
SA(R=500)

Figure 3.2


Percent of optimal allele number
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0
20
40
60
80
100
Core Set size

Chapter 4

Choosing Tagging SNPs based on Entropy

Kejun Liu[1], Li Li[1], and Spencer Muse[1,2]

1. Bioinformatics Research Center, North Carolina State University

   Raleigh, NC 27695

2. Correspondence:

   Dr Spencer V Muse

   501 Partners II, Bioinformatics Research Center

   North Carolina State University

   Raleigh, NC 27695

   Phone: (919) 515-1948

   Fax: (919) 515-7315

   Email: muse@stat.ncsu.edu

ABSTRACT

Recent findings suggest that the human genome can be divided into discrete blocks in a way such that the variation of each block can be captured by a small number of SNPs (referred to as tagging SNPs). Here we describe a new, efficient two-stage method to choose tagging SNPs based on pairwise linkage disequilibrium (LD) and haplotype entropy. The first stage uses a numerical algorithm for recursively partitioning the chromosomal region into several blocks, with the expectation that much of the recombination occurred between blocks. The second stage consists of two different local search methods to choose tagging SNPs based on entropy. For blocks with a large numbers of SNPs an incremental search was implemented to obtain a practically useful solution, while a decremental search was used to obtain an arbitrarily chosen level of quality when the block is tractable for haplotype estimation. We investigated the efficiency of our algorithms using population simulations based on both homogeneous recombination and hotspot recombination models. Our main results can be summarized as follows: 1) Pairwise linkage disequilibrium provides a reliable but conservative way for block partitioning. The quality of the partition depends on the background recombination rate and hotspot recombination rate for a given chromosomal region. 2) Haplotype entropy is demonstrated analytically and by simulation to be a good measure for choosing tagging SNPs. Our algorithms provide efficient and accurate identification of tagging SNPs.

INTRODUCTION

Evidence has been given that the human genome may be divided into blocks separated by recombination hotspots (Daly *et al.* 2001; Patil *et al.* 2001; Gabriel *et al.* 2002; Dawson *et al.* 2002). Recent attention has focused on the incorporation of block structure to map common disease genes (Johnson *et al.* 2002; Zhang *et al.* 2002b). Such studies employ a haplotype map of single nucleotide polymorphisms (SNPs) to detect association between a haplotype block and disease. Construction of haplotype maps is currently underway. Biologically, a haplotype block has been defined as a contiguous chromosomal region with little or no evidence of historical recombination (Gabriel *et al.* 2002). A major limitation to this concept has been lack of knowledge of the recombination rate of the human genome, which is widely believed to be highly non-homogeneous over distances less than 100 kb (Kruglyak 1999). There are two general classes of methods to define a block based on genetic information content. One is based on haplotype diversity, and the other is based on pairwise linkage disequilibrium (LD). In a recent paper, Patil *et al.* (2001) defined a region of consecutive SNPs as a block if at least α percent of the haplotypes in the block are represented more than once in a sample in the absence of missing data. The concept of ambiguous and unambiguous haplotypes was introduced when missing data are present. The philosophy behind this is that both haplotype number and diversity are limited in the absence of recurrent mutation and recombination. The other approach to a working definition of blocks, which relies on linkage disequilibrium, assumes a strong correlation between high LD and low recombination rate. For example, Gabriel *et al.* (2002) defined a haplotype block as a region with a small proportion of comparisons among SNP pairs showing strong evidence of historical recombination.

Because LD is also dependent on the age of the mutations (i.e. allele frequency) and on evolutionary forces such as population structure and demographic history, it is necessary to allow for some low levels of pairwise linkage disequilibria within a block.

A key question facing any block partitioning algorithm is: how likely does a "block" incorrectly include a recombination hotspot? The answer depends on local patterns of recombination across the block sequence, but also relies on our definition of blocks. The greedy algorithm developed by Patil *et al.* (2001), later implemented as a dynamic programming algorithm by Zhang *et al.* (2002a), aims to minimize the number of SNPs required to distinguish α percent of the unambiguous haplotypes in the blocks. Wiuf *et al.* (2003) pointed out that their algorithms may fail to identify obvious recombination hotspots when hotspots are inferred from SNP patterns. Alternatively, Gabriel *et al.* (2002) proposed a block partitioning algorithm for unphased genotype data using D' confidence intervals. Although the results from this study are convincing, the method appears to be specific to their particular data. Here we present a new definition of a haplotype block based on genotype data, and an efficient algorithm to perform the block partitioning to maximize the possibility of recombination hotspot identification. We use population simulations with hotspot recombination models to validate our algorithms. Although our algorithm still uses a cutoff value like other methods, the algorithm is designed so that when a more stringent cutoff value is used, new blocks are formed by breaking down existing blocks.

In each block, a large proportion of common haplotypes can be distinguished by a small subset of "tagging" SNPs (Clayton 2001; Johnson 2001). These tagging SNPs can potentially be useful for association studies, in that much genotyping effort can be avoided and the multiple testing problems can be simplified. It is also argued that if the disease association of a specific allele depends on a specific haplotype, the disease association may not be detected when each SNP of the haplotype is analyzed individually. Efforts have been made to select tagging SNPs from both short and long chromosomal regions. The two general classes of methods to partition blocks are also used to choose tagging SNPs. Meng *et al.* (2003) proposed a spectral decomposition method for selecting markers based on composite pairwise linkage disequilibrium, not assuming Hardy-Weinberg equilibrium. This method does not depend on a block structure and targets a large number of SNPs. When the haplotype frequencies are available, either determined experimentally or estimated statistically, different measures are used to compare different sets of tagging SNPs based on haplotype frequencies, and exhaustive searches are generally employed to search all possible subsets to obtain the "best" set of tagging SNPs. Clayton *et al.* (2001) proposed the percentage of haplotype diversity explained (PDE) as the measure of haplotype diversity, suggesting the exhaustive search as the method to search the solution space with the added constraint of a maximal number of five tagging SNPs. Others (Gabriel *et al.* 2002; Ke and Cardon 2003) implicitly or explicitly use the coverage value (see method: haplotype tagging) as the criterion. In both cases, the exhaustive search will be computationally infeasible for large numbers of SNPs.

We provide justifications and efficient algorithms for selecting tagging SNPs based on haplotype entropy (Shannon 1948). We formally derive the relationship between entropy and tagging SNPs. The use of entropy as a measure of haplotype diversity was also mentioned in Ackerman *et al.* (2003) and Judson *et al.* (2002). In contrast to their studies, we do not require the haplotype frequencies for the whole block to be estimated, as the estimates might be both inaccurate and computationally infeasible when the possible haplotype number is extremely large. When the number of SNPs in the block lies in the range where haplotype estimation can be accurately estimated from the population data (e.g. SNP number <10), the decremental search provides an efficient way to automatically find the set of tagging SNPs. Again, we use population simulations to validate our algorithms.

METHODS

Block Partitioning

Assume that we are given N diploid individuals with genotypes of n consecutive SNPs known. We wish to partition the chromosomal sequence into discrete blocks with the maximal value of R, where R denotes the ratio of between-block recombination rate to within-block recombination rate. In the case of the hotspot model described in Wiuf and Posada (2003), the algorithm aims to exclude all the recombination hotspots from any block and assign the region between two consecutive hotspots as a single block, which is the scenario of obtaining the maximal value of R. Consequently, blocks partitioned by this criterion represent regions with low levels of recombination, whereas intervals between blocks represent the opposite.

A block can be defined as a chromosomal region where recombination hotspots do not exist and within-block background recombination is weak enough so that the correlation between pairwise linkage disequilibrium and physical proximity is not significant. A working definition that captures the essence of this description declares a sequence of SNPs to be a block if any bisection of this sequence does not significantly improve the within-block pairwise linkage disequilibrium. If a bisection can significantly improve within-block LD, the sequence will be recursively bisected until each sub-sequence is a block. We formulate the definition and the algorithm as follows. Let $r_i, i = 1, 2..n$ be the *i*th SNP locus and define $d_{ij}, i \neq j$, as the measure of linkage disequilibrium between the *i*th and *j*th locus. For convenience we define $d_{ii} = 1$ for $i = 1..n$. We do not require a

specific measure of linkage disequilibrium, but we do require that d_{ij} lies in the range (0..1), with the value of 1 representing the maximal linkage disequilibrium. In this paper $|D'|$ and r^2 are used as pairwise linkage disequilibrium measures. For a sequence of n loci, the ith bisection partitions the sequence into two sub-sequences $(r_1,...,r_i)$ and $(r_{i+1},...,r_n)$. We define statistics T_i for the ith bisection as the relative increase of within-block pairwise LD resulting from the ith bisection. Formally,

$$b_i = \frac{\sum_{j=1}^{i}\sum_{k=i+1}^{n} d_{jk}}{i*(n-i)}$$

$$w_i = \frac{\sum_{\substack{j,k=1 \\ j<k}}^{i} d_{jk} + \sum_{\substack{j,k=i+1 \\ j<k}}^{n} d_{jk}}{i*(i-1)/2 + (n-i)*(n-i-1)/2}$$

$$T_i = \frac{w_i - b_i}{b_i}$$

where w_i is simply the weighted average LD within the two sub-sequences, the value b_i is the average LD between the two sub-sequences. The relative difference of these two values, T_i, can be interpreted as follows: if the interval between the ith locus and the $(i+1)$th locus is a recombination hotspot, or if the whole region tends to show larger LD values for physically linked loci pairs, then T_i will be significantly larger than 0. The "best" bisection can be found by choosing the interval with the maximal T_i (denoted as T). A simple partitioning algorithm can be developed based on T: for the current sequence we compute T. If $T > c$, where c is a cutoff selected by the user, we bisect the sequence into two sub-sequences in the interval providing the value of T (if there are multiple intervals with the same value of T we simply choose the first one). Otherwise,

we regard the sequence as a block. If a bisection occurred, the same algorithm is applied to each one of the sub-sequences. The procedure is repeated until all subsequences are defined as blocks. The result of the algorithm is inclusive for different significant values, in the sense that a small value of c will conserve all the block boundaries found using higher values of c. In other words, blocks defined using a small c value were either inherited or generated by breaking down blocks that would be defined using higher c values.

We measure the quality of a partitioning in two ways. First we define the identification index (I) as the probability of recombination hotspots being assigned to intervals between two contiguous blocks. We call a block partitioning is *reliable* if $I > 0.95$. A block partitioning method is suspicious if its identification index is small. The second quality, the accuracy (A) of the block portioning, is the normalized relative ratio of recombination rates (defined as $R / \max(R)$, where $\max(R)$ is the maximal value that R can obtain). Partitioning methods that are *reliable* and $A > 0.80$ are said to be *accurate*.

SNP Selection based on Entropy

After partitioning the chromosomal region into several discrete blocks, for each block we wish to select the minimal subset of SNPs needed to capture (most of) the within-block haplotype variation. Given a block containing r SNPs and N phased diploid individuals, the following measure of haplotype diversity is defined on any subset of SNPs (S)

$$E_S = -\sum_{i=1}^{m_S} \tilde{P}_i \log_2(\tilde{P}_i),$$

where m_S is the haplotype number for subset S and $\tilde{P}_i$ is the ith observed haplotype frequency. Thus, E_S is just the sample haplotype entropy for the SNP subset S. If E_{all} is the sample haplotype entropy for the whole set of SNPs, then it is not difficult to show that $E_S \le E_{all}$ for any subset of any size (Appendix).

We define a set of SNPs S as tagging SNPs if $E_S = E_{all}$ is satisfied. If the coverage of the subset (C_S) is defined as the maximal proportion of the whole-set haplotypes that can be unambiguously distinguished by SNPs in subset S, then tagging SNPs also satisfy $C_S = 1$. The reverse is also true. Interestingly, tagging SNPs can equivalently be defined as follows: a subset S is a set of tagging SNPs if $E_S = E_{S'}$ holds for all S', where S' is formed by all the SNPs in S and any additional SNP not in S. The proof can be found in Appendix. This definition motivates an efficient method to test whether a subset is a set of tagging SNPs without knowing the haplotype information for the whole set, and provides a formal justification for the use of entropy for identifying tagging SNPs.

In practice, phase is unknown or only partially known for genotype data. An estimated haplotype entropy of the given subset can be defined as

$$\hat{E}_S = -\sum_{i=1}^{m_S} \hat{P}_i(S) \log_2(\hat{P}_i(S)),$$

where $\hat{P}_i(S)$ is the estimated haplotype frequency. Any appropriate statistical methods can be used to estimate the haplotype frequencies. In this work haplotype estimation is performed by the Expectation-Maximization (EM) algorithm (Excoffier and Slatkin,

1995). We assume that for estimated haplotype entropy the relationship of entropy and tagging SNPs still approximately holds, with the definition of tagging SNPs changing to the subset of SNPs that explains a large proportion (>95%) of haplotype entropy (E_{all}). In other words, E_S / E_{all} drops from 1 to .95 to account for the uncertainty introduced by having to estimate the haplotype frequencies.

Suppose a block contains r SNPs, for small values of r (e.g, < 10) the following decremental search was developed to find the minimal subset of tagging SNPs:

1) Set $i = r$. Estimate the haplotype frequency for all the r SNPs. Denote the estimated haplotype entropy as $\hat{E}_{all}$. Accept the whole set as the initial candidate set of tagging SNPs (denoted as $S_{best}(i)$, where i is the size of the set).
2) For each subset $S(i-1) \in S_{best}(i)$, the subhaplotypes of size i (corresponding to $S_{best}(i)$) were pooled into subhaplotypes of size $i-1$ (corresponding to $S(i-1)$) as follows: if any subhaplotype of size i shares the same alleles over all SNPs of $S(i-1)$, they should be pooled into a single subhaplotype of size $i-1$. Haplotype entropy for $S(i-1)$ was evaluated. If the maximal haplotype entropy of all possible subsets ($\hat{E}_{best}(i-1)$) is not significantly smaller than $\hat{E}_{all}$, the corresponding subset $S_{best}(i-1)$ is accepted as the new candidate of tagging SNPs. Otherwise, the algorithm breaks and return the last candidate of tagging SNPs as the solution.
3) Repeat step 2) by setting $i = i-1$ until $i = 2$. Return the last candidate of tagging SNPs as the solution.

When r increases linearly, the possible haplotype number increases exponentially, making haplotype estimation both inaccurate and computationally infeasible. For large values of r, we developed the following incremental search method:

1) Set $i = 2$. For each of the $\binom{r}{2}$ subsets of size 2, estimate the haplotype frequencies and compute the estimated haplotype entropy. Retain the subset $S_{best}(2)$ with the maximal entropy value $\hat{E}_{best}(2)$ as the initial candidate set of tagging SNPs.
2) For each of the SNPs that are not included in the last candidate set, evaluate the haplotype entropy of all subsets of size $i+1$ by adding one unselected SNP into the candidate subset, then accept the SNP bringing the maximal entropy increase in the new candidate. Denote the new candidate set as $S_{best}(i+1)$
3) The candidate subset from 2) is optimized by the following iterative search. All $S(i) \in S_{best}(i+1)$ are evaluated, and the subset ($S'_{best}(i)$) having the maximal estimated entropy is retained. From the remnant set $\{j; j \notin S'_{best}(i)\}$, add the SNP bringing the maximal entropy increase (denote the SNP as j') and set $S_{best}(i+1) = S'_{best}(i) + \{j'\}$. The step of swapping is repeated until the procedure converges. Set the converged subset $S_{best}(i+1)$ as the new candidate set of tagging SNPs with entropy value $\hat{E}_{best}(i+1)$. If $\hat{E}_{best}(i+1)$ is significantly larger than $\hat{E}_{best}(i)$, $S_{best}(i+1)$ is accepted as the new candidate set of tagging SNPs. Otherwise, the algorithm breaks and return $S_{best}(i)$.

4) Repeat step 2) and step 3) by setting $i = i + 1$ until $i = r$. Return the whole set as the solution.

For all the simulations we assume $\hat{E}_a$ is not significantly smaller than $\hat{E}_{all}$ if $\hat{E}_a \geq 0.95\hat{E}_{all}$, or $\hat{E}_a$ is significantly larger than $\hat{E}_b$ if $\frac{\hat{E}_a - \hat{E}_b}{\hat{E}_b} \geq 0.01$.

The Coalescence Process with Recombination Hotspots

We carried out simulation of $2N$ haplotypes consisting of n consecutive SNPs in two successive steps. In the first step, a random genealogy of the entire sample of $2N$ haplotypes was generated by running the simulation backwards in time and keeping track of common ancestor and recombination events (Hudson 1983; Hudson and Kaplan 1995). Once the ancestral relationships of haplotypes were generated, mutations were then placed on the genealogy using an infinite-allele model. The infinite-allele model assumes that each segregating locus was created by one and only one mutation event in the population history. At each locus, the mutation was placed on a single branch leading to a node with the number of descendants lying in the range of $0.5N$ and $1.5N$, with the probability of selecting a branch proportional to its length. Consequently, the frequency of marker allele was constrained to lie in the range 0.25-0.75. This constraint reflects the fact that in practice, only SNPs with appreciable polymorphism levels are used.

The coalescence process with hotspot recombination was implemented using the procedure described in Wiuf and Posada (2003). This model is an extension of the

coalescent with uniform recombination rate based on the idea that recombination break points are concentrated in certain regions of the chromosomes. We assume that all the hotspots have the same recombination rates, and that the imprecision of the recombination hotspots is 0, implying that the recombination hotspots center on single regions between two surrounding loci. We also assume the hotspot sites uniformly distribute across the chromosome. Denote R_b as the per generation background recombination rate at any non-hotspot interval between two consecutive loci, and define R_h as the hotspot recombination rate for any hotspot interval. Let k be the hotspot number, n be the total SNP number, then the global recombination rate is $R_g = R_b * (n-k-1) + R_h * k$. Note that R_b and R_h are properties of the chromosomal region, whereas R is a quantitative measure depending on results of the block partitioning algorithm. Under this model, the maximal value of R is R_h / R_b.

For all the simulations we assumed a constant effective population size of 30,000. We assume that all the SNPs are distributed evenly spaced across the genome, and the physical distance between any two consecutive SNPs is 1kb. Once the $2N$ haplotypes were generated, N diploid individuals were formed by random association among those haplotypes.

RESULTS

Block Partitioning

Coalescent simulations were carried out to generate 23 populations of 200 haplotypes. For each one of the 100 physically evenly spaced SNPs, allele frequencies were constrained to lie between 0.25-0.75 (other choices show similar results). Hotspot number varied from 1 to 5, with the same probability for each value. The background recombination rate (R_b) and the hotspot recombination rate (R_h) varied in the 23 simulated populations. R_b varied from 10^{-8} to 10^{-5}, and R_h varied from 10^{-5} to 10^{-3}. The settings of recombination rates were based on the observation that the human genome-wide recombination rate is approximately 10^{-5} per kb. For each population, 100 individuals were first generated by random associations among the 200 simulated haplotypes; the block partitioning algorithm was then applied to the genotype data assuming phase was unknown. For each simulated population, 6 different cutoff values were used to estimate the block structure. 100 independent replicates were performed for each setting of parameters. Table 4.1 presents the identification index (I) and accuracy (A) estimated from the true hotspot information recorded in the simulation step, using $|D'|$ as the linkage disequilibrium measure. I was estimated by the average percentage hotspots excluded from estimated blocks. A was first estimated by the average ratio of the among-block recombination rate to the within-block recombination rate in the estimated block structure using the assigned recombination rates, then normalized by $R_{\max}$. We see in this table that I primarily depends on r_h and c, but not on r_b. When the hotspot recombination rate is not significantly larger than the genome-wide average, it is

difficult to identify the hotspots by our algorithm, even for extremely low background recombination rates. A hotspot with ten-fold intensity (compared to genome-wide average) produces reliable blocks for small cutoff values, but not for large cutoff values. For more intense hotspots, which might be common for the human genome, our algorithms have the potential to identify almost all the hotspots ($I \approx 1$). Although our algorithm seems to exclude hotspots from blocks, the comparisons of estimated R values with the maximal R values (A) suggest our algorithm is conservative. For example, when r_h is ten-fold higher compared to genome average ($r_h = 10^{-4}$), the best A values for different cutoffs range from 0.28 to 0.87 under different number of hotspots, showing a low accuracy of the block partitions for high background recombination rates. When r_h becomes higher and r_b is at least 10 times smaller than genome average, block partitioning becomes reasonably accurate and shows $A > 0.80$ for certain cutoff values, a condition of block partitioning we call *accurate*. The most accurate partitioning is found in small cutoff values when r_h is small and in large values when r_h is large. This observation is expected. When two cutoff values both show a reliable hotspot identification, high cutoff value will show a better accuracy since small cutoff value may partition the chromosomal region into small blocks, a major factor driving R values away from the optimal value. However, when r_h is small, high cutoff values will fail to identify all the hotspots therefore will not generate good blocks. In practice we suggest multiple values of c be examined until the majority of the blocks show a significant strong within-block LD and/or the number of the SNPs in the block is appropriate for further study.

To assess the influence of LD measure on the reliability and accuracy of the algorithm, we use r^2 as pairwise LD measure and repeat the simulation process. Table 4.2 shows the result. Two observations emerge from this table. First, r^2 outperforms $|D'|$ for large background recombination rates but $|D'|$ performs slightly better for low r_b values. This may reflect the quality of these two measures under different recombination rates. When background recombination rate is near to 0, $|D'| \approx 1$ is approximately independent of other factors whereas r^2 is still dependent on allele frequencies. However, when r_b is large, r^2 may provide a more reasonable measure than $|D'|$. It is well known that $|D'|$ tends to inflate for small allele frequencies (Gabriel *et al.* 2002). Second, the best cutoff values of r^2 is larger than $|D'|$. For example, in table 4.2 the best cutoff for $r_b = 10^{-5}, r_h = 10^{-3}$ is 5, whereas the best cutoff is 1 if $|D'|$ is the LD measure. This may result from the fact that on average $|D'|$ values are larger than r^2 values, and the scale of the average LD values in the denominator of T_1 dominates the value of the statistics.

Haplotype Tagging

We first explored the range of entropy values for a block under different settings of recombination rates and SNP numbers. We simply assumed a homogeneous background recombination rate within a block. For a specific number of SNPs per block (r) and per generation recombination rate (R_b) between two consecutive SNPs, a sample size ($2N$) of 200 and 1000 haplotypes were generated separately using the coalescence process and the infinite-site mutation model. This procedure was repeated 100 times for each setting of parameters, to obtain the average entropy values in table 4.3. For most cases entropy

increases as R_b and/or r increases. For small R_b s the increase of entropy with r is slower than that of large R_b s. For example, when $2N$ = 200, as k increases from 1 to 15, the entropy value increases from 0.93 to 6.58(or 7-fold) when $R_b = 10^{-4}$, but increases only to 1.78(or two-fold) when $R_b = 10^{-7}$. When R_b is small, entropy is approximately independent of the sample size, whereas entropy increases monotonically as sample size increases for large R_b s. Table 4.3 also indicates that a sample size of 200 haplotypes is sufficient to capture >95% of haplotype variation measured (in terms of haplotype entropy) in 1000 haplotypes when $R_b \leq 10^{-5}$. Note that haplotype entropy also relies on SNP frequency.

A similar simulation was carried out to verify that the identified tagging SNPs have the correct nominal entropy level (>95%) under different conditions. Shown in table 4.4 are the average percent of entropy obtained (E_S / E_{all}) and average coverage (C_S) from 100 replicates. The entropy and coverage were computed from the true haplotype frequencies, and S here represents the estimated set of tagging SNPs. For each simulated population both incremental and decremental searches were performed. This table clearly shows both methods are reliable except for $R_b = 10^{-4}$, where the decremental search performs well but the incremental search tends to underestimate the tagging SNP number. For all cases the decremental search shows $E_S / E_{all} > 0.95$, even though S was evaluated from estimated entropies. There is a strong correlation between E_S / E_{all} and C_S. For values near 1, these two measures can be regarded as equivalent.

Figure 4.1 plots the average number of tagging SNPs with the total number of SNPs for the decremental search. The incremental search shows a very similar pattern. When $R_b = 10^{-7}$, a block with 15 SNPs can be tagged with 3.1 SNPs, on average. This number increases to 3.8 and 6.8 when R_b is set to 10^{-6} and 10^{-5}, respectively. When $R_b = 10^{-4}$, at least 9 SNPs are required to explain >95% of the haplotype variation (in terms of haplotype entropy).

DISCUSSION

We argued that before choosing tagging SNPs, the chromosomal region being studied should first be partitioned into blocks of consecutive SNPs, within which linkage disequilibrium (both two-locus LD and multiple-locus LD) is strong. Such a partitioning could be based on any block partitioning algorithm, knowledge of the recombination rates for different regions, or both. In this paper we developed a block partitioning algorithm to maximize the possibility of hotspot identification. Our algorithm reveals strong hotspots under a less-stringent definition of blocks, and continues to reveal existing hotspots and also discloses less strong hotspots when a more stringent block definition is used. The degree of hotspot identification is controlled by a single cutoff value. We verified our block partitioning definition and algorithm using population simulations. The simulation results show that recombination hotspots can be reliably identified, whereas the "true" hotspots are mixed with "false" hotspots when the background recombination plays a role in the realization of current LD patterns. Therefore, our algorithm not only detects intervals that are recombination hotspots, but also breaks a large chromosomal region with homogeneous recombination rates into several blocks.

To find out if the phase information has a strong effect on the block partitioning algorithm, we also performed the partitioning algorithm on the simulated populations assuming phase is known. A slightly faster performance is observed due to the computational time saved in the two-locus haplotype frequency estimation procedure. However, we did not see any significant increase in qualities of block partitioning over

all the combinations of parameters (data not shown). The estimated blocks are identical most of the time. We observed that two-locus linkage disequilibrium can be accurately estimated from unphased genotype data when Hardy-Weinberg equilibrium holds. Therefore, little value was added from the phase information in our pairwise-LD based partitioning algorithm.

A variety of statistics have been proposed to measure the haplotype variation and identify the set of tagging SNPs (Johnson et al. 2001; Ke and Cardon 2003; Ackerman *et al.* 2003; Judson *et al.* 2003). Compared to these studies, our methods have several features. First, we pointed out that haplotype entropy, which measures the haplotype variation, has a theoretical relationship with tagging SNPs. A perfect set of tagging SNPs explains all the haplotype variation and has the same information (entropy) as the whole set of SNPs. A practically useful set of tagging SNPs explains the majority of the haplotype variation and includes most of the information (entropy). Second, previous studies for choosing tagging SNPs were challenged by numerical complexity, as exhaustive searches were used to find the minimum SNP subset. As a result of the exponential nature of the exhaustive search, there exists a computational limit on all of these methods. For example, the online supplement to the paper by Johnson et al. (2001) provides executable programs limiting the maximum subset size to 5 (we found 5 SNPs are not sufficient to capture the haplotype variation in many cases). Our algorithms for choosing tagging SNPs take the efficiency into account. By avoiding the exhaustive search of all possible subsets of SNPs for large blocks (for small blocks exhaustive search can be used), our algorithm has the potential to converge to local optima, therefore should be regarded as an approximate

method. However, our algorithms select the tagging SNPs capturing the nominal haplotype variation efficiently, with equal or slightly larger tagging SNP numbers compared to those from an exhaustive search. In the case of an association study, a conservative approach is preferred. Finally, we provide a practically useful approach for choosing tagging SNPs from very large blocks (e.g. SNP number > 20) for unphased genotype data. In this case, computational haplotype inference for the whole set of SNPs, which is required for all previous methods, is either inaccurate or computationally infeasible.

The knowledge of these blocks and sets of tagging SNPs will put research in a stance to carry out association studies with more informative multiple-locus haplotypes than less informative bi-allelic SNPs. In contrast to single marker based approaches, which look for differences in frequencies of a single SNP between affected individuals and controls, haplotype-based approaches look for differences in frequencies of haplotypes between the populations. Such studies have been shown to be very useful (Akey *et al.* 2000; Johnson *et al.* 2001; Zhang *et al.* 2002). The two-stage strategy developed in this paper makes the haplotype-based approach applicable to unphased genotype data. The optimal procedure for performing haplotype-based tests, as well as the effect of this strategy on the power of association studies, is currently under investigation.

APPENDIX: Relationship of entropy and tagging SNPs

We are given a block with r SNPs and m unique haplotypes. Define $\tilde{P}_i, i = 1..m$, as the observed frequency of the ith haplotype h_i, the sample entropy for the whole set of SNPs $E_{all} = -\sum_{i=1}^{m} \tilde{P}_i . \log_2(\tilde{P}_i)$. S is a subset of $\{1, 2..., n\}$, $|S| = r_S$. If only SNPs in subset S were considered, then some of the original haplotypes might become indistinguishable and will be pooled. Suppose the subset S grouped the original haplotypes into m_S unique subhaplotypes $h_i(S), i = 1..m_S, m_S \le m$, then the frequency of the ith subhaplotype $\tilde{P}_i(S) = \sum_{i'=1}^{m} \tilde{P}_{i'} . b_{i'}(i; S)$, where $b_{i'}(i; S) = 1$ if the allelic state in $h_{i'}$ is identical to that of $h_i(S)$ at all SNP loci in subset S, otherwise $b_{i'}(i; S) = 0$. $E_S = -\sum_{i=1}^{m_S} \tilde{P}_i(S) . \log_2(\tilde{P}_i(S))$ is defined as the haplotype entropy of subset S.

We define S as a set of tagging SNPs if $m_S = m$ holds. This indicates that the same number and same frequencies of haplotypes would be generated if only the tagging SNPs are genotyped. Thus, the entropy of the set of tagging SNPs will be the same as the entropy of the whole set. On the other hand, if a subset S satisfies $E_S = E_{all}$, then S must be a set of tagging SNPs. Here we describe a simple proof. For the ith unique subhaplotype $h_i(S)$, there exists a set of $m_i (m_i \ge 1)$ different original haplotypes(denote the set as Δ) satisfying $b_{i' \in \Delta}(i; S) = 1$. Denote $\tilde{P}_{i'}(i; S), i' = 1..m_i$ as the the original (observed) frequency of i'th haplotype in Δ, then $\tilde{P}_i(S) = \sum_{i'=1}^{m_i} \tilde{P}_{i'}(i; S)$. It can easily be

shown that the entropy contribution for each subhaplotype $h_i(S)$ is always less than or equal to the summation of the entropy contribution for the original haplotypes $r_1(i;S), r_2(i;S), ..., r_{m_i}(i;S)$:

$$-\tilde{P}_i(S).\log_2(\tilde{P}_i(S)) = -(\sum_{i'=1}^{m_i}\tilde{P}_{i'}(i;S)).\log_2(\sum_{i'=1}^{m_i}\tilde{P}_{i'}(i;S)) \le -\sum_{i'=1}^{m_i}(\tilde{P}_{i'}(i;S).\log_2(\tilde{P}_{i'}(i;S)))$$

where the equality holds if and only if $m_i = 1$. Then, each subhaplotype $h_i(S)$ has a unique original haplotype h_j if and only if the equality holds for all $i, i = 1..m_S$. In other words, $m_S = m$ holds if and only if $E_S = E_{all}$.

A more general argument, which can obviously be extended by the above proof, is that $E_S \le E_{S'}$ if S is a subset of S'. The equality holds if and only if the map between subhaplotypes of S' and subhaplotypes of S is one-to-one.

Define $\overline{S} = S_{all} - S, S_{all} = \{1, 2, ..., r\}$, $\overline{S}_l$ as the lth SNP of $\overline{S}$, $S_l^{'} = S + \{\overline{S}_l\}$, then $H_S \le H_{S_l^{'}}$ for all $l = 1..r - r_S$, and the equality holds if and only if the allelic state of $\overline{S}_l$ is unambiguously determined by the allelic state of the SNPs in S for each subhaplotype $r_i(S_l^{'}), i = 1..m_S$. If $H_S = H_{S_l^{'}}$ for all $l = 1..r - r_S$, each subhaplotype $h_i(S)$ will unambiguously map to a unique haplotype $h_{i'}$; thus we obtain $m_S = m$, and S is a set of tagging SNPs.

In practice, we replace sample entropy E with estimated entropy $\hat{E}$ and we compare the difference between $\hat{E}_{best}(i)$ with $\hat{E}_{best}(r)$ or $\hat{E}_{best}(i+1)$, where $\hat{E}_{best}(i)$ denotes the largest entropy in all possible subsets satisfying $r_S = i$. A subset of all SNPs S is called an empirical set of tagging SNPs if $\hat{E}_S \geq 0.95\hat{E}_{all}$. We are interested in finding the empirical set of tagging SNPs with the minimal size.

ACKNOWLEDGEMENTS

I would like to express my gratitude to David Posada, who provided the hotspot simulation manuscript and source code.

REFERENCES

Ackerman H, Usen S, Mott R, Richardson A, Sisay-Joof F, Katundu P, Taylor T, Ward R, Molyneux M, Pinder M and Kwiatkowski D, Haplotypic analysis of the TNF locus by association efficiency and entropy, *Genome Biology* **4**: R24 (2003).

Akey J, Jin L, Xiong M, Haplotypes vs single marker linkage disequilibrium tests; what do we gain? *European J of hum Genet* **9**: 291-300 (2000)

Clayton D, Choosing a set of haplotype tagging SNPs from a larger set of diallelic loci. ftp-gene.cimr.cam.ac.uk/software (2001).

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES. High-resolution haplotype structure in the human genome. *Nature Genet* **29**: 229-232 (2001).

Dawson E, Abecasis GR, Bumpstead S, Chen Y, Hunt S, Beare DM, Pabial J, A first-generation linkage disequilibrium map of human chromosome 22. *Nature* **418**: 544-548 (2002).

Dickson, LE, A New Solution of the Cubic Equation, *Amer. Math.* Monthly **5**, 38-39, (1898).

Excoffier L, Slatkin M, Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* **12**: 921-927 (1995).

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D, The structure of haplotype blocks in the human genome. *Science* **226**: 225-2229 (2002).

Hudson RR, Properties of a neutral allele model with intragenic recombination. *Theor. Pop. Biol.* **23**: 183-201 (1983).

Hudson RR and Kaplan N, Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**: 147-164 (1990).

Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eavaes IA, Dudbridge F, Twells RC, Payne F, Hughes W, Nutland S, Stevens H, Carr P, Tuomilehto-Wolf E, Tuomilehto J, Gough SC, Clayton DG, Todd JA, Haplotype tagging for the identification of common disease genes. *Nat Genet* **29**: 233-237 (2001).

Judson R, Salisbury B, Schneider J, Windemuth A and Stephens JC, How many SNPs does a genome-wide haplotype map require? *Pharmacogenomics* **3**: 279-391 (2002).

Ke X and Cardon LR, Efficient selective screening of haplotype tag SNPs, *Bioinformatics* **19**, 287-288 (2003).

Kruglyak L, Prospects for whole-genome linkage disequilibrium mapping of common disease genes, *Nat Genet* **22**, 139-144 (1999).

Liu K and Muse SV, PowerMarker: new genetic data analysis software. Version 1.0. Free program distributed by the author over the internet from http://www.powermarker.net

Meng Z, an electronic thesis submitted to North Carolina State University. Available online at http://etd.ncsu.edu (2003)

Nordborg M and Tavare S, Linkage disequilibrium: what history has to tell us. *TRENDS in Genetics*, **18**: 83-90 (2002).

Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, Nguyen BT, Norris MC, Sheehan JB, Shen N,

Stern D, Stokowski RP, Thomas DJ, Trulson MO, Vyas KR, Frazer KA, Fodor SP, Cox DR, Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**: 1719-1723 (2001).

Shannon CE, A Mathematical Theory of Communication, *The Bell System Technical J.* **27**, 379-423 (1948).

Weir BS, Genetic Data Analysis II. Sunderland, MA: Sinauer (1996).

Wiuf Carsten and Posada David, A coalescent model of recombination hotspots. To appear in *Genetics* (2003)

Zhang K, Deng M, Chen T, Waterman MS and Sun F, A dynamic programming algorithm for haplotype block partitioning. *Proc Natl Acad Sci* USA, **99**: 7335-7339 (2002a)

Zhang K, Calabrese P, Nordborg M and Sun F, Haplotype block structure and its applications to association studies: Power and study designs. *Am. J. Hum. Genet.* **71**: 1386-1394 (2002b).

Table 4.1: Performance of recursive bisection algorithm based on $|D'|$

Background recombination rate (R_b)	Cutoff (c)	Hotspot recombination rate (R_b)					
		1×10^{-5}	1×10^{-4}	1×10^{-3}	1×10^{-5}	1×10^{-4}	1×10^{-3}
		I			A		
1×10^{-8}	0.1	0.32	0.98	1.00	0.11*	0.87*	0.97
	0.25	0.18	0.94	1.00	0.06	0.79	0.98*
	0.5	0.09	0.83	1.00	0.03	0.58	0.98
	1	0.03	0.61	1.00	0.01	0.31	0.98
	2	0.01	0.27	0.97	0.00	0.10	0.90
	5	0.00	0.05	0.50	0.00	0.02	0.33
1×10^{-7}	0.1	0.32	0.99	1.00	0.19*	0.79*	0.89
	0.25	0.18	0.95	1.00	0.13	0.77	0.94
	0.5	0.09	0.85	1.00	0.09	0.61	0.97*
	1	0.04	0.61	1.00	0.04	0.32	0.97
	2	0.00	0.28	0.96	0.01	0.12	0.89
	5	0.00	0.03	0.47	0.01	0.02	0.31
1×10^{-6}	0.1	0.38	0.98	1.00	0.21	0.35	0.38
	0.25	0.24	0.94	1.00	0.22*	0.45	0.54
	0.5	0.12	0.85	1.00	0.20	0.52*	0.71
	1	0.03	0.62	1.00	0.13	0.46	0.87*
	2	0.00	0.20	0.95	0.10	0.18	0.87
	5	0.00	0.00	0.30	0.10	0.01	0.21
1×10^{-5}	0.1	-	0.97	1.00	-	0.17	0.09
	0.25	-	0.92	1.00	-	0.19	0.13
	0.5	-	0.81	1.00	-	0.23	0.20
	1	-	0.38	0.99	-	0.28*	0.47*
	2	-	0.00	0.34	-	0.10	0.30
	5	-	0.00	0.00	-	0.10	0.01

* Maximal accuracy for different cutoff values.

Results are averaged over 100 replicates. I and R were calculated from the estimated block partitioning and the settings of recombination rates.

Table 4.2: Performance of recursive bisection algorithm based on r^2

Background recombination rate (R_b)	Cutoff (c)	Hotspot recombination rate (R_b)					
		1×10^{-5}	1×10^{-4}	1×10^{-3}	1×10^{-5}	1×10^{-4}	1×10^{-3}
		I			R		
1×10^{-8}	0.1	0.61	1.00	1.00	0.11	0.34	0.35
	0.25	0.45	0.99	1.00	0.12*	0.58	0.59
	0.5	0.30	0.98	1.00	0.08	0.73	0.77
	1	0.19	0.94	1.00	0.06	0.76*	0.89
	2	0.08	0.86	1.00	0.02	0.65	0.96
	5	0.02	0.58	1.00	0.01	0.30	0.99*
1×10^{-7}	0.1	0.68	1.00	1.00	0.14	0.25	0.26
	0.25	0.51	1.00	1.00	0.19	0.48	0.50
	0.5	0.36	0.99	1.00	0.20*	0.65	0.70
	1	0.21	0.96	1.00	0.15	0.73*	0.85
	2	0.10	0.86	1.00	0.09	0.63	0.93
	5	0.02	0.56	1.00	0.03	0.31	0.98*
1×10^{-6}	0.1	0.68	1.00	1.00	0.15	0.12	0.11
	0.25	0.52	1.00	1.00	0.18	0.19	0.19
	0.5	0.36	0.99	1.00	0.20	0.31	0.32
	1	0.22	0.96	1.00	0.22*	0.47	0.51
	2	0.11	0.83	1.00	0.21	0.53*	0.73
	5	0.03	0.44	1.00	0.14	0.34	0.95*
1×10^{-5}	0.1	-	0.99	1.00	-	0.14	0.05
	0.25	-	0.99	1.00	-	0.15	0.07
	0.5	-	0.98	1.00	-	0.17	0.09
	1	-	0.92	1.00	-	0.20	0.14
	2	-	0.75	1.00	-	0.28*	0.28
	5	-	0.15	0.95	-	0.25	0.76*

* Maximal accuracy for different cutoff values.

Results are averaged over 100 replicates. I and R were calculated from the estimated block partitioning and the settings of recombination rates.

Table 4.3: Entropy values for different settings

No. of SNPs [a] (r)	Entropy values							
	Within-block recombination rate (R_b)							
	10^{-7}		10^{-6}		10^{-5}		10^{-4}	
	2N=200	*2N=1000*	*2N=200*	*2N=1000*	*2N=200*	*2N=1000*	*2N=200*	*2N=1000*
1	0.93	0.93	0.92	0.93	0.93	0.94	0.94	0.94
2	1.20	1.20	1.20	1.22	1.46	1.50	1.78	1.80
3	1.40	1.37	1.35	1.43	1.76	1.79	2.54	2.59
4	1.40	1.47	1.52	1.51	2.13	2.18	3.31	3.33
5	1.52	1.56	1.60	1.61	2.37	2.47	3.92	4.03
6	1.54	1.53	1.75	1.70	2.64	2.68	4.49	4.64
7	1.63	1.58	1.87	1.88	2.86	2.93	4.96	5.14
8	1.61	1.66	1.84	1.89	3.07	3.12	5.36	5.61
9	1.65	1.70	1.95	1.93	3.23	3.26	5.65	6.00
10	1.76	1.73	2.06	2.04	3.39	3.48	5.90	6.35
11	1.70	1.74	2.07	2.09	3.54	3.67	6.10	6.61
12	1.78	1.68	2.13	2.12	3.73	3.80	6.25	6.85
13	1.73	1.77	2.16	2.19	3.79	3.92	6.39	7.05
14	1.78	1.79	2.18	2.21	3.91	4.04	6.53	7.23
15	1.78	1.81	2.24	2.29	4.13	4.18	6.58	7.34

Results are averaged over 100 replicates. Entropy values were calculated from haplotype frequencies.

[a] Number of SNPs per block

Table 4.4: Obtained entropy level and coverage

No. of SNPs [a] (r)	Method[b]	Within block recombination rates(R_b)							
		10^{-7}		10^{-6}		10^{-5}		10^{-4}	
		E_S / E_{all}	C_S	E_S / E_{all}	C_S	E_S / E_{all}	C_S	E_S / E_{all}	C_S
3	D	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00
	I	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00
4	D	1.00	1.00	1.00	1.00	0.99	0.99	1.00	1.00
	I	1.00	1.00	1.00	1.00	0.99	0.99	1.00	1.00
5	D	1.00	1.00	0.99	1.00	0.98	0.99	1.00	1.00
	I	1.00	1.00	0.99	1.00	0.98	0.99	1.00	1.00
6	D	1.00	1.00	0.99	1.00	0.98	0.99	1.00	1.00
	I	1.00	1.00	0.99	1.00	0.98	0.98	0.99	0.99
7	D	0.99	1.00	0.99	1.00	0.98	0.98	0.98	0.97
	I	1.00	1.00	0.99	1.00	0.97	0.98	0.98	0.97
8	D	0.99	1.00	0.99	0.99	0.97	0.98	0.97	0.94
	I	0.99	1.00	0.99	0.99	0.96	0.96	0.96	0.93
9	D	0.99	1.00	0.99	0.99	0.97	0.97	0.97	0.95
	I	0.99	1.00	0.98	0.99	0.96	0.96	0.94	0.88
10	D	0.99	1.00	0.98	0.99	0.97	0.97	0.97	0.93
	I	0.99	1.00	0.98	0.99	0.95	0.95	0.93	0.85
11	D	0.99	1.00	0.98	0.99	0.97	0.97	0.97	0.92
	I	0.99	1.00	0.98	0.99	0.94	0.94	0.91	0.82
12	D	0.99	1.00	0.98	0.99	0.97	0.97	0.96	0.92
	I	0.99	1.00	0.98	0.99	0.94	0.93	0.91	0.80
13	D	0.99	1.00	0.98	0.99	0.97	0.96	0.96	0.91
	I	0.99	1.00	0.98	0.99	0.93	0.93	0.90	0.78
14	D	0.99	1.00	0.98	0.99	0.97	0.96	0.96	0.91
	I	0.99	1.00	0.98	0.99	0.93	0.92	0.89	0.76
15	D	0.99	0.99	0.98	0.99	0.97	0.96	0.96	0.91
	I	0.99	0.99	0.98	0.99	0.93	0.91	0.89	0.75

Results are averaged over 100 replicates.

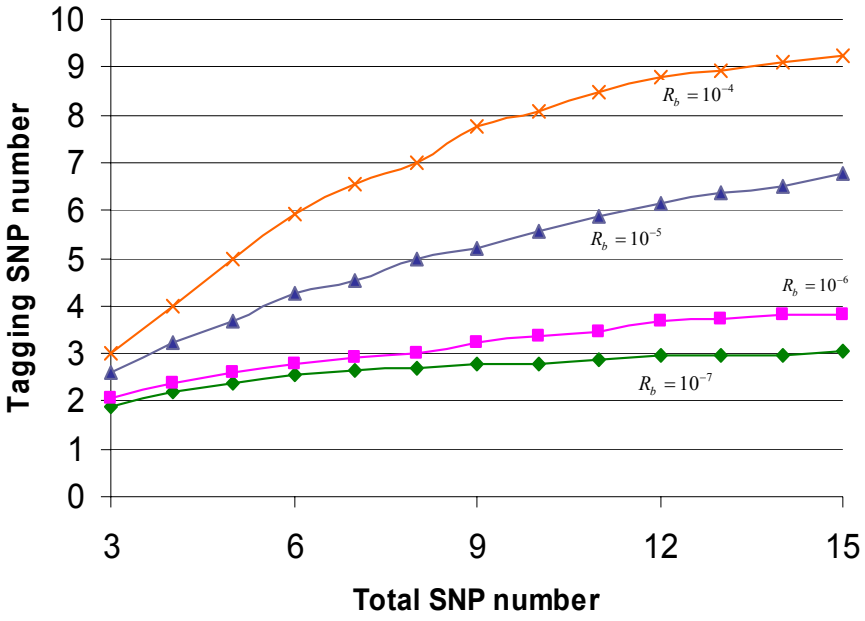
[a] Number of SNPs per block

[b] Method used to search the minimal set of tagging SNPs. D: decremental search; I: incremental search.

Figure Legends

Figure 4.1: Relationship between average number of tagging SNPs and number of all SNPs for different recombination rates. A decremental search was performed for each of the 100 replicates.

Figure 4.1


Tagging SNP number
Total SNP number
0
1
2
3
4
5
6
7
8
9
10
3
6
9
12
15
$R_b = 10^{-4}$
$R_b = 10^{-5}$
$R_b = 10^{-6}$
$R_b = 10^{-7}$

Chapter 5

PowerMarker Package

INTRODUCTION

With genetic markers becoming basic tools for geneticists, the need for reliable computer software and libraries to perform statistical/computational analysis of marker data has grown. One of the main reasons that we have developed the PowerMarker package is to satisfy this need for elegant, but simple, reusable solutions. PowerMarker delivers a data-driven, integrated analysis environment (IAE) for marker data. The IAE integrates the data management, analysis and visualization in a user-friendly graphic user interface. It accelerates the analysis lifecycle, and enables users to maintain data integrity throughout the process.

Analyses implemented in PowerMarker can be organized into six different categories. Summary analysis computes summary statistics for each marker or taxon, and offers modules to choose the optimal subset of markers or lines. The recursive bisection partitioning algorithm, described in chapter 4, is also included in this group. Estimation analysis estimates allelic, genotypic and haplotypic frequencies from unrelated data or family data. Disequilibrium analysis calculates Hardy-Weinberg and two-locus linkage disequilibrium statistics for quantifying or testing the disequilibrium between two alleles at the same locus, and at different loci, respectively. Structure analysis estimates classical F-statistics and population specific F-statistics, and constructs pairwise coancestry matrices. Phylogenetic analysis computes allele frequencies for each taxon and calculates pairwise distances based on the frequencies. Phylogenetics trees, representing the phylogentic relationship among different taxa, can also be reconstructed. Association analysis performs tests for detecting a significant association between a trait and single

marker or a specific haplotype. Apart from the different types of analyses, PowerMarker also provides a coalescence simulation module for generating a large amount of data for experimental analysis.

PowerMarker offers a comprehensive library of reusable classes for genetical marker data analysis, especially for SSR/SNP data analysis, PowerMarker can be accessed by programmers through any Microsoft .NET language such as C++ and Visual Basic. The PowerMarker package, as well as the full documentation of PowerMarker interface and source codes, will be delivered to the academic community without charge. This chapter covers a brief tutorial to the graphic interface, and lists the methods implemented in PowerMarker.

TUTORIAL

This tutorial is designed to demonstrate the graphic interface of PowerMarker. Following the steps in this tutorial will allow the user to learn about using PowerMarker's integrated analysis environment to perform analysis. This tutorial shows how to:

- Create a project for the analyses
- Import a DataSet from text file
- Choose a subset from the DataSet
- Produce a table of summary statistics
- Compute linkage disequilibrium coefficients and view the results in PowerMarker

Launch the PowerMarker application by double-clicking its icon  on the desktop to begin the tutorial.

Step 1: Creating the project

Before performing an analysis in PowerMarker, you must first create a project to work in. PowerMarker uses a project file with a **.prj** extension to organize data objects and folders. If this is your first time to run PowerMarker, you will notice a project Default is automatically created and displayed in the explorer like this:

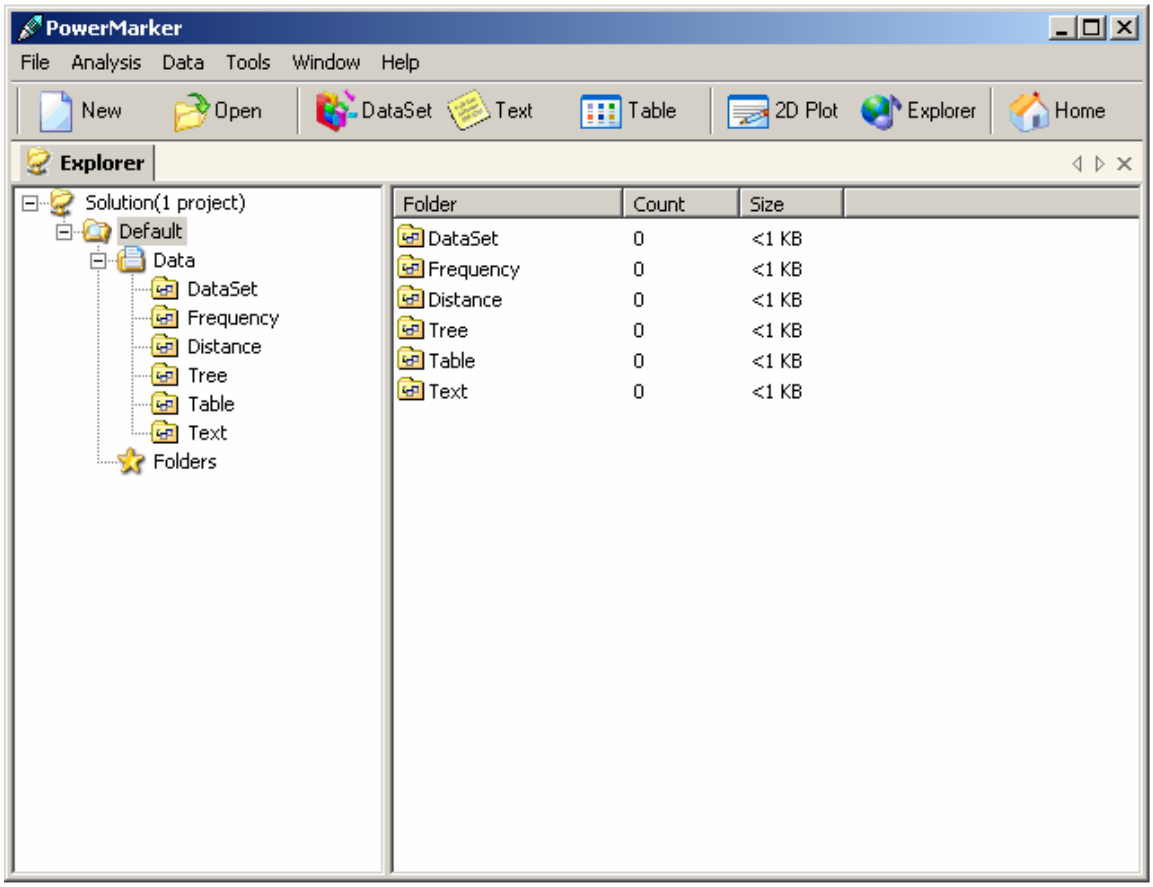


Figure 5.1: The object explorer in PowerMarker

The following steps can be used to create your own project:

1. Choose **File | Close all projects** to close all projects.
2. Choose **File | Add New Project** or click the **New** button on the main toolbar to open the file dialog to save the new project.
   - Accept the default directory name, and type **fbi** in the file name field.
   - Click the **Save** button to close the dialog.
3. Now the interface changes back to Figure 5.1 except that the name of the project name has been changed to **fbi**.

Step 2: Importing a DataSet

The majority of the analyses in PowerMarker works on a DataSet. A DataSet is a serialized object of genetic marker data. To import a DataSet from a text file, follow these steps:

1. Choose **File | Import | DataSet** or click the **DataSet** button on the main toolbar to open the DataSet wizard.
   - Click **Browse** button to open the file dialog
   - Choose the file **fbi.txt** from **\<PowerMarker>\Samples\FBI**, where **<PowerMarker>** is the directory where you installed PowerMarker.

   Step 1 of the DataSet wizard should look like this:

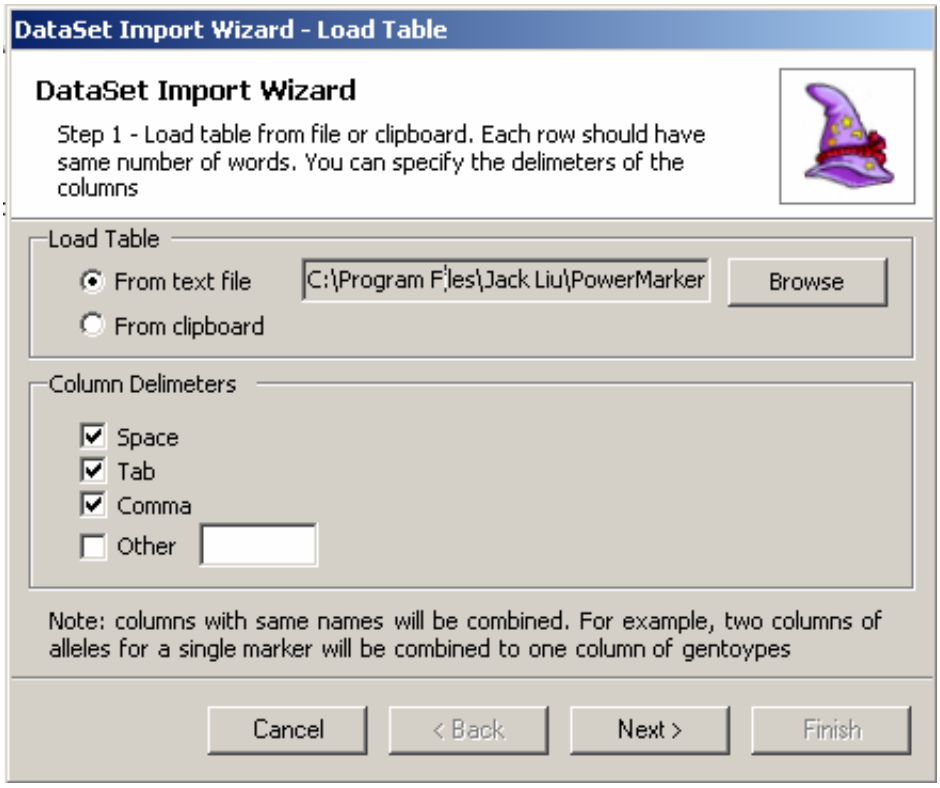


Figure 5.2: Step 1 of Data wizard

2. Click the **Next** button to go to step 2 of the DataSet wizard.

   - Select the first two columns **Sample** and **ID#,** and click the link **Categorical** to change these two columns to categorical types. All the other columns are accepted as marker types.
   - Select **ID#** from the drop down list of Level-1 Column combobox.
   - Select **Sample** from the drop down list of Level-2 Column combobox.

   Step 2 of the DataSet wizard should look like this:

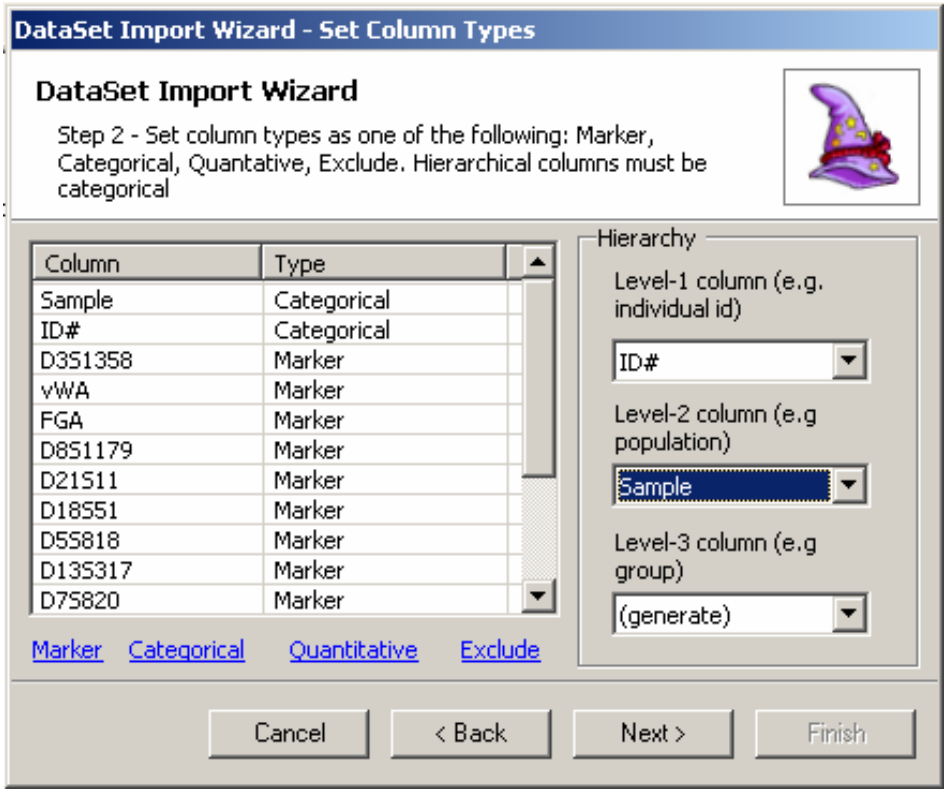


Figure 5.3: Step 2 of Data wizard

3. Click the **Next** button to go to step 3 of the DataSet wizard. Accept all the settings in step 3.

4. Click the **Next** button to go to step 4 of the DataSet wizard. Step 4 of the wizard should like this:

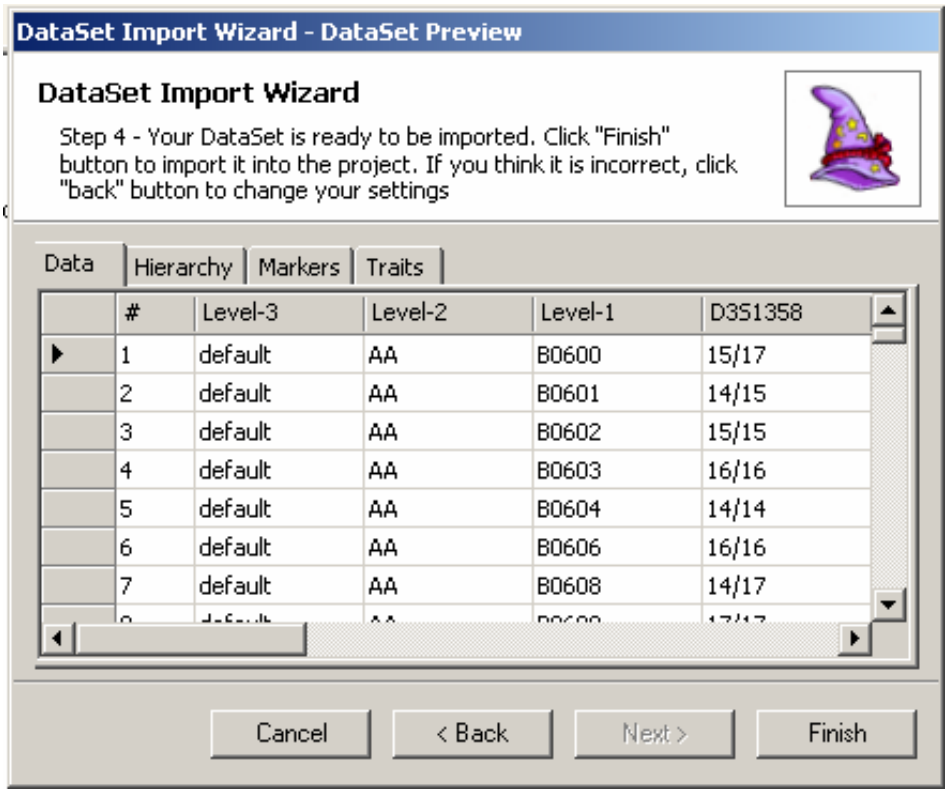


Figure 5.4: Step 4 of Data wizard

5. Click the **Finish** button to go back to the explorer.

You will notice that the DataSet *fbi* has been imported to the project.

Step 3: Choosing a subset from the DataSet

This step will generate a new DataSet from *fbi* by excluding loci with a large missing proportion (>0.05).

1. Right click the newly generated DataSet *fbi* and select **Choose subset** from the pop-up menu. The choose subset dialog will appear.

2. Switch to the **Choose markers** tab. By default all the markers are selected.
    - Select **Missing proportion** from the drop down list of the combobox.
    - Click the **Compute property** button.
    - Click the header **Missing proportion** in the ListView
    - Select the first 6 markers.

    The dialog should look like this:

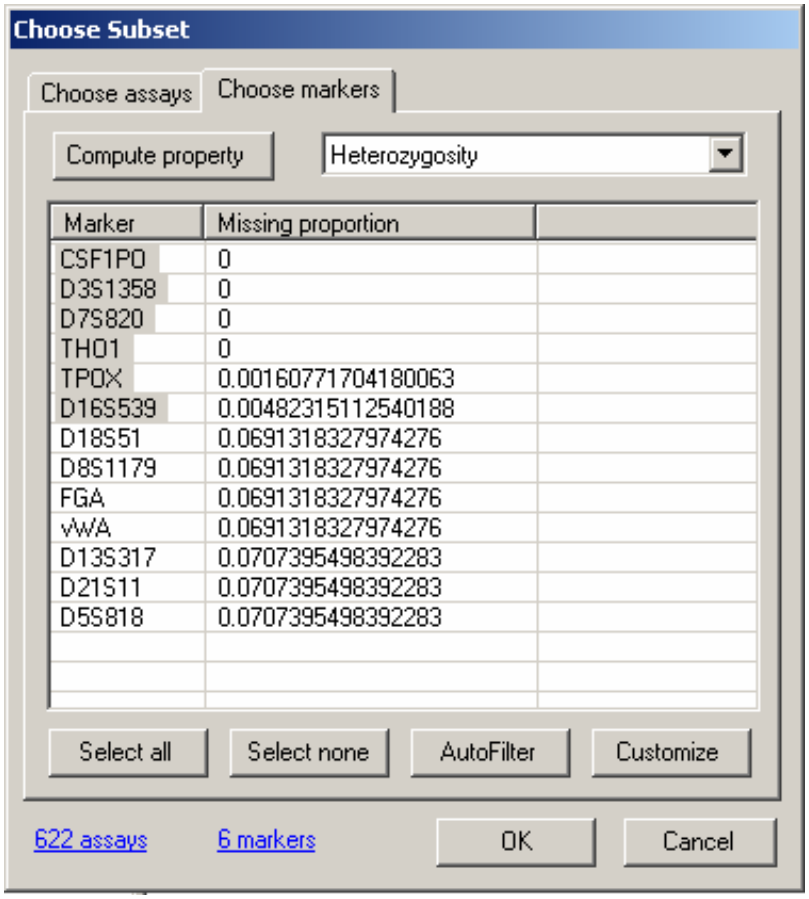


Figure 5.5: Choose subset dialog

3. Click **OK** button to close the dialog. A new DataSet *fbi.SubData* will appear in the explorer.

Step 4: Producing a table of summary statistics

In this step we will generate a table of summary statistics for the DataSet *fbi.SubData*.

1. Choose **Analysis | Summary | Summary Statistics** to open the analysis dialog. Make the following changes in the appropriate fields in the dialog:

   - Select **fbi.SubData** in the Data ListBox.
   - Type **Summary** as the name of the result folder.

   The dialog should like this:

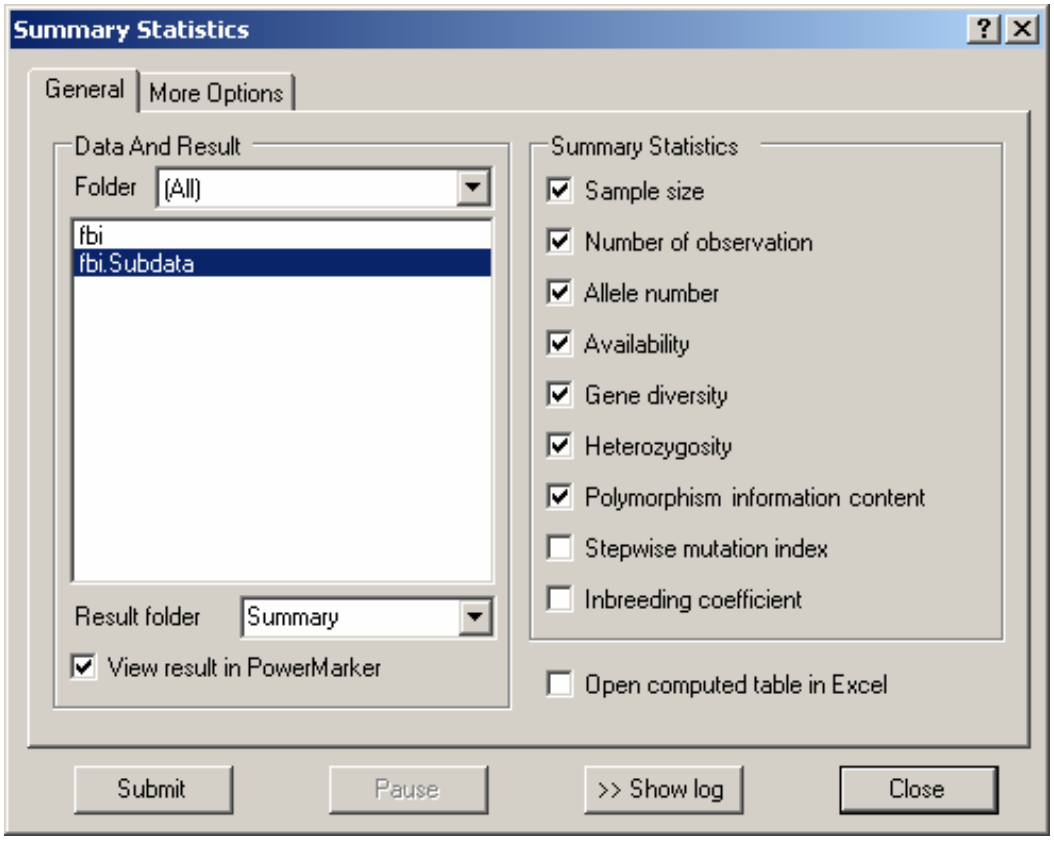
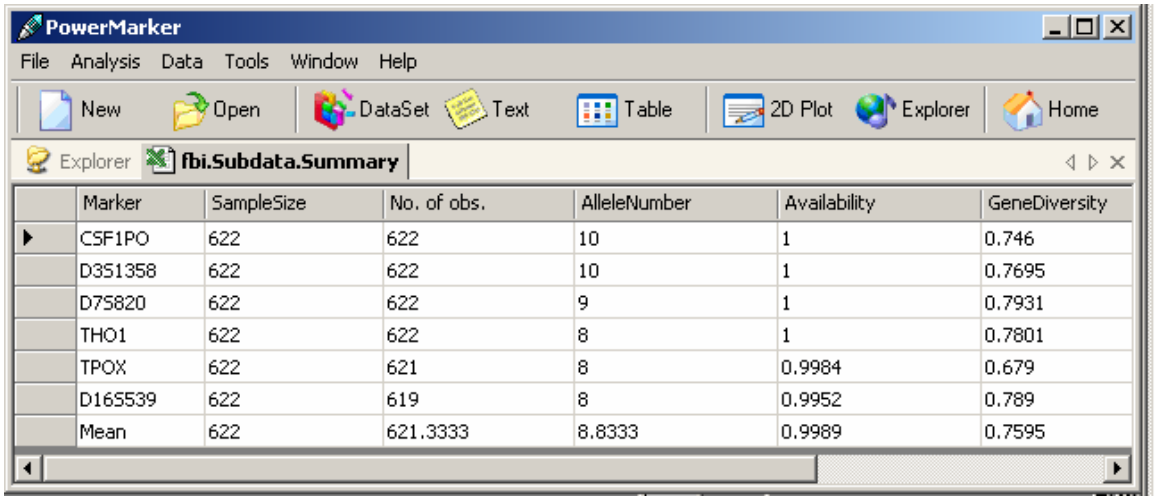


Figure 5.5: Analysis dialog for summary statistics

2. Click **Submit** button to perform the analysis. The analysis should be finished immediately. PowerMarker will automatically save the result and open it in TableViewer.



Marker	SampleSize	No. of obs.	AlleleNumber	Availability	GeneDiversity
CSF1PO	622	622	10	1	0.746
D3S1358	622	622	10	1	0.7695
D7S820	622	622	9	1	0.7931
THO1	622	622	8	1	0.7801
TPOX	622	621	8	0.9984	0.679
D16S539	622	619	8	0.9952	0.789
Mean	622	621.3333	8.8333	0.9989	0.7595

Figure 5.7: Table viewer in PowerMarker

3. Right click in the TableViewer and select **Open In Excel** from the pop-up menu. The table will be opened in Excel.

Step 5: Computing linkage disequilibrium coefficients and Viewing the results

In this step we will generate a matrix of linkage disequilibrium statistics for the DataSet fbi.

1. Choose **Analysis | Disequilibrium | Two-Locus Linkage Disequilibrium 2D Matrix** to open the analysis dialog. Make the following changes in the appropriate fields in the dialog:
   - Select **fbi** in the Data ListBox.
   - Type **LD** as the name of the result folder.
   - Choose **D'** as the upper triangle statistics, **ChiSquare** as diagonal statistics and **(none)** as lower triangle statistics.

The dialog should like this:

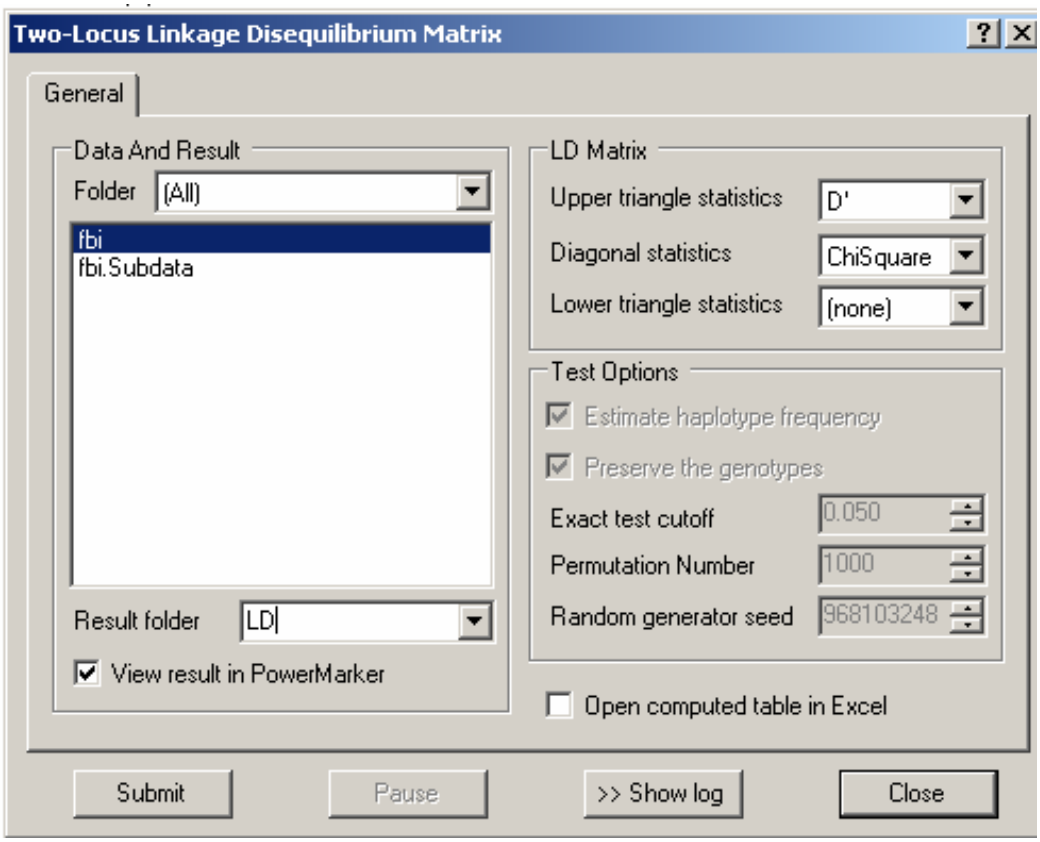


Figure 5.8: Analysis dialog for Two-Locus linkage disequilibrium

2. Click **Submit** button to perform the analysis. After the analysis is finished, PowerMarker will automatically save the result and open it in TableViewer. The result table is named as *fbi.ldmatrix*.
3. Right click table *fbi.ldmatrix* and choose **2D Plot** from the popup menu. 2D viewer will be opened. Initially all cells will be blank (white).
4. Click the **Add** button on the left to open the range dialog. Make changes to appropriate fields:
   - Click all points <= , and type **0.05** in the field.
   - Check **Apply to diagonal**, uncheck the other two options.

- Click **Choose color** link to open the color picker, select the blank color and close the picker.

The dialog should look like this:

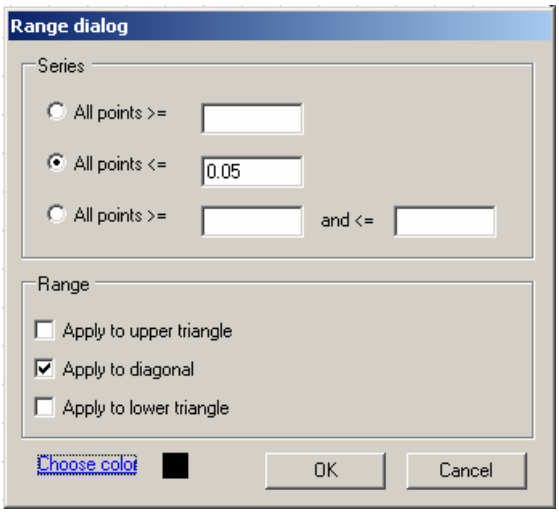


Figure 5.9: Range dialog

- Click **OK** button to close the dialog. A new series will appear in the **series** listbox.
- Click **Draw series** link to draw the series. The interface should look like this:

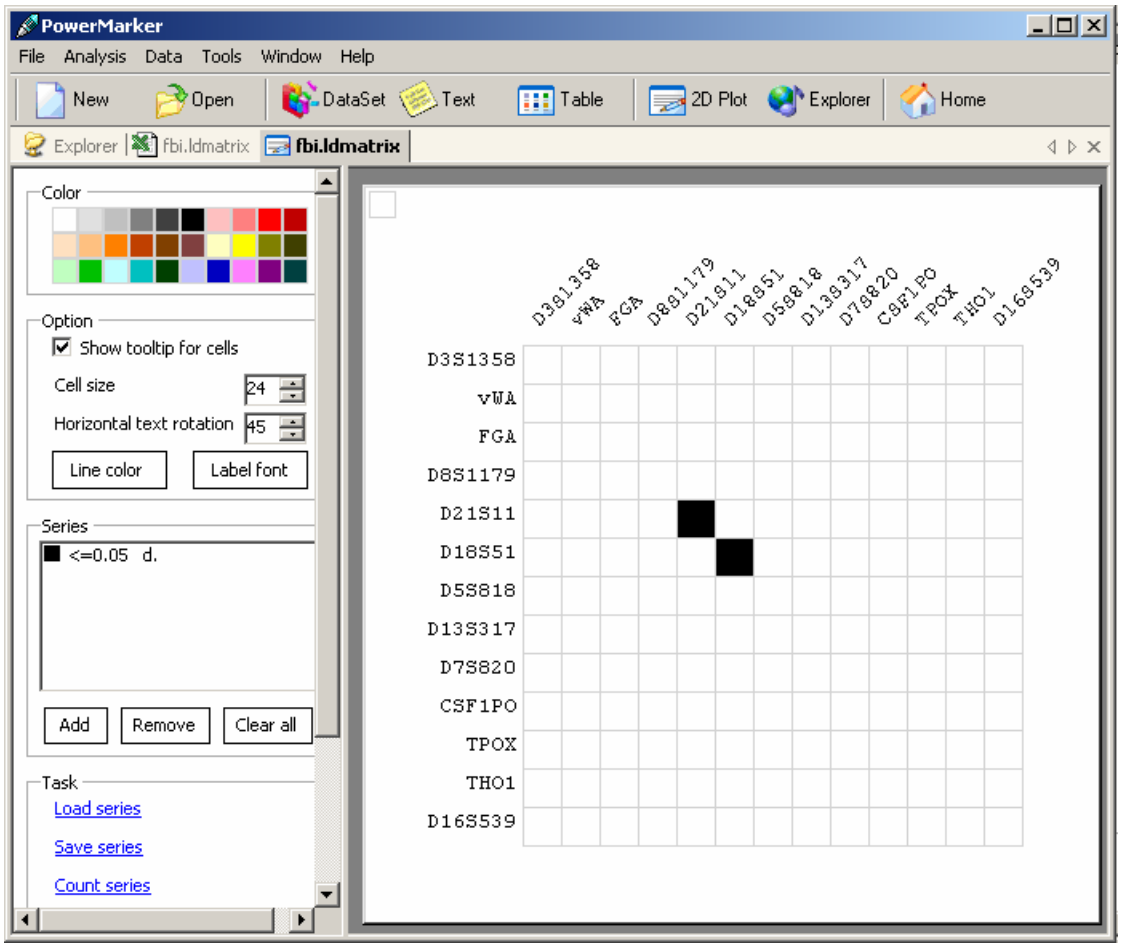


Figure 5.10: 2D plot

METHODS

For a more detailed description of the equations and methods in this chapter, see the cited pages from Weir (1996) unless otherwise noted. When possible, we use the notation and concepts in Weir (1996).

Suppose we are given n individuals and m polymorphic loci. The symbol A will be used to mean any genetic locus with a series of alleles A_u. For an individual, a single-locus genotype or a single allele is observed for each locus. An allele A_u has a population frequency p_u (or p_{lu}, to indicate the lth locus), and a genotype $A_u A_v$ has a population frequency P_{uv} (or P_{luv}),. Sample frequencies will be indicated by tildes, and these observed values are also used as estimates of allelic and genotypic frequencies. Estimates will be indicated by carets. In a sample, counts of alleles and genotypes will be written as n_u and n_{uv} (or n_{lu} and n_{luv} for the lth locus), respectively.

Summary

Basic statistics

The number of observation for a marker locus is defined as the number of nonmissing alleles (for haploid data) or nonmissing genotypes (for diploid data) observed in the sample. A genotype is missing if one of its two alleles is missing. Availability is defined as:

$$1 - \frac{Obs}{n},$$

where Obs is the number of observations and n is the number of individuals sampled.

Stepwise mutation index is defined as the maximal proportion of alleles that follows the stepwise mutation pattern.

Within-population inbreeding coefficient

An EM algorithm (pp. 77-78) is used to find the MLE of the within-population inbreeding coefficient. Note that the EM algorithm may fail to converge for negative values of inbreeding coefficient. The same parameter is estimated using the method of moments (pp. 79-80).

Diversity indices

Heterozygosity (pp. 141-150) is simply the proportion of heterozygous individuals in the population. At a single locus it is estimated as

$$\hat{H}_l = 1 - \sum_{u=1}^{k} \tilde{P}_{luu}$$

Gene diversity (pp. 150-156), often referred to as expected heterozygosity, is defined as the probability that two randomly chosen alleles from the population are different. An unbiased estimator of gene diversity at the *l*th locus is

$$\hat{D}_l = (1 - \sum_{u=1}^{k} \tilde{p}_{lu}{}^2)/(1 - \frac{1+f}{n}),$$

where the inbreeding coefficient, f, is estimated from the data using the method of moments (pp. 79-80). The user can also request the common biased estimator of the gene diversity,

$$\widehat{D_l} = (1 - \sum_{u=1}^{k} \tilde{p}_{lu}^2).$$

A closely related diversity measure is the polymorphism information content (PIC) (Botstein *et al.* 1980). It is estimated as

$$\widehat{PIC}_l = 1 - \sum_{u=1}^{k} \tilde{p}_{lu}{}^2 - \sum_{u=1}^{k-1} \sum_{v=u+1}^{k} 2\tilde{p}_{lu}^2 \tilde{p}_{lv}^2$$

For all of these diversity measures, the overall estimates are calculated as the average across all loci, whereas variances and confidence intervals are estimated by nonparametric bootstrapping across different loci.

CoreSet

The CoreSet module selects the optimal set of individuals to capture the maximal variation in the sample for a given core set size. A batch script system is developed for supporting user-defined constraints and settings (see chapter 3).

Estimation

Allele and genotype frequencies

The sample allele frequencies are calculated as $\tilde{p}_u = n_u /(2n)$, with the variance estimated as

$$\text{var}(\tilde{p}_u) \hat{=} \frac{1}{2n}(\tilde{p}_u + \tilde{P}_{uu} - 2\tilde{p}_u^2),$$

where $\hat{=}$ means "estimated by".

The sample genotype frequencies $\tilde{P}_{uv}$ are calculated as n_{uv}/n. Both the $\tilde{p}_u$ s and $\tilde{P}_{uv}$ s are unbiased maximum likelihood estimates (MLEs) of the population frequencies. Confidence intervals for allele and genotype frequencies are formed by resampling individuals from the data set.

Haplotype estimation

The EM algorithm (Excoffier and Slatkin 1995) was implemented to estimate haplotype frequencies and to probabilistically assign phases for genotypes. The multiple-tier

optimization of the implementation, as well as the extended algorithm for pedigree incorporation, is described in appendix C.

Disequilibrium

Hardy-Weinberg disequilibrium

For a single locus, the MLE of the disequilibrium coefficient D_{uv} for alleles A_u and A_v is calculated as

$$\hat{D}_{uv} = \begin{cases} \tilde{P}_{uv} - \tilde{p}_u \tilde{p}_v, & u = v \\ \tilde{p}_u \tilde{p}_v - \frac{1}{2}\tilde{P}_{uv}, & u \neq v \end{cases},$$

and the variance is estimated using the follow formulas:

$$\mathrm{Var}(\hat{D}_{uu}) \triangleq \frac{1}{n}\left[\tilde{p}_u^2 (1-\tilde{p}_u)^2 + (1-2\tilde{p}_u)^2 \hat{D}_{uu} - \hat{D}_{uu}^2 \right]$$

$$\mathrm{Var}(\hat{D}_{uv}) \triangleq \frac{1}{2n}\left\{ \tilde{p}_u \tilde{p}_v (1-\tilde{p}_u)(1-\tilde{p}_v) + \sum_{w \neq u,v} \left(\tilde{p}_u^2 \hat{D}_{uw} + \tilde{p}_v^2 \hat{D}_{vw} \right) \right.$$
$$\left. -\left[(1-\tilde{p}_u - \tilde{p}_v)^2 - 2(\tilde{p}_u - \tilde{p}_v)^2 \right] \hat{D}_{uv} + \tilde{p}_u^2 \tilde{p}_v^2 - 2\hat{D}_{uv}^2 \right\}.$$

Bootstrap confidence intervals are formed by resampling individuals from the data set. Three different methods are used to test for Hardy-Weinberg Equilibrium. The chi-square goodness-of-fit test is formed by calculating the chi-square statistic

$$X_T^2 = \sum_u \frac{(n_{uu} - n\tilde{p}_u^2)^2}{n\tilde{p}_u^2} + \sum_u \sum_{v \neq u} \frac{(n_{uv} - 2n\tilde{p}_u \tilde{p}_v)^2}{2n\tilde{p}_u \tilde{p}_v} \ .$$

This statistic has $k(k-1)/2$ degrees of freedom where k is the number of alleles at the marker locus. The same distribution is shared by the likelihood ratio test described in Weir (pp. 105-106). A permutation version of the exact test given by Guo and Thompson (1992) is also implemented (pp. 109-100).

Two-Locus Linkage disequilibrium

Two-locus linkage disequilibrium D_{uv} is defined for two alleles at different loci as: $D_{uv} = p_{uv} - p_u p_v$. It is estimated by $\hat{D}_{uv} = \tilde{p}_{uv} - \tilde{p}_u \tilde{p}_v$ for haplotype data or phased genotype data, or by $\hat{D}_{uv} = \hat{p}_{uv} - \tilde{p}_u \tilde{p}_v$ for unphased genotype data. The analytic solution for $\hat{p}_{uv}$ is given in chapter 4. Based on the estimates of D_{uv}, five linkage disequilibrium measures are calculated for each pair of alleles at two loci: the correlation coefficient r^2, Lewontin's D', the proportional difference d, the population attributable risk δ, and Yule's Q. These measures are discussed in Devlin and Risch (1995).

The chi-square statistic to test that all the pairwise linkage disequilibrium D_{uv} are zero is calculated as follows:

$$X_T^2 = \sum_{u=1}^{k} \sum_{v=1}^{l} \frac{(2n)\hat{D}_{uv}^{\ 2}}{\tilde{p}_u \tilde{p}_v}.$$

This statistic has $(k-1)(l-1)$ degrees of freedom for markers with k and l alleles, respectively. Permutation versions of the exact test for testing whether two-locus genotype frequencies are the products of one-locus frequencies (not assuming HWE), or testing whether two-locus genotypic frequencies are products of allele frequencies (assuming HWE), are also implemented. The details of these methods can be found in Weir (pp. 127-128).

Multi-Locus linkage disequilibrium

The exact test for multi-locus association, described by Zaykin *et al.* (1995), is implemented.

Structure

Classic F-Statistics

PowerMarker performs 4 different types of F-statistics analysis. The first type works on haploid data or diploid data (assuming HWE), and reports the overall estimate $\hat{\theta}$ and an estimator for each locus. The details are given in Weir (pp. 170-174). When analyses are performed at the genotypic level, the same general approach is followed except three levels of F-statistics are now estimated (pp. 176-179): F indicates the degree of inbreeding within individuals, θ is the inbreeding coefficient between alleles of different individuals, and f measures the degree of inbreeding within populations. A three-level hierarchical analysis, described by Weir (pp. 184-186), is also implemented under the optional assumption of Hardy-Weinberg Equilibrium.

Population specific F-Statistics

The procedure of estimating population specific F-Statistics and between-population F-Statistics was formulated in Weir and Hill (2002). An extension of the estimation procedure, which works on genotype frequencies instead of allele frequencies, can be found in Appendix A.

Coancestry matrix

A coancestry matrix is formed by calculating θ for each pair of populations. The user can request for the log transformation ($= -\ln(1-\theta)$) to be performed.

Phylogeny

Frequency-based distance

Various distance measures used for frequency data have been described by Nei (1987) and Weir (1996). Appendix B lists the definitions and brief explanations of the distance measures implemented in the package.

Tree reconstruction and bootstrap

The following two algorithms are used to reconstruct the phylogeny from a distance matrix: UPGMA(unweighted pair-group method using arithmetic average) and Neighbor-joining (pp. 344-356). Bootstrapping is performed over the marker loci (Felsenstein 1985). Each bootstrap sample consists of same number of markers sampled with replacement from the original data set, and it then is subjected to the same distance calculation and tree reconstruction. The output is a list of trees that can be summarized to obtain a consensus tree by the program "consensus" in Phylip package (Felsenstein 1993).

Association test

Case control test

PowerMarker offers three methods for testing an association between a single marker and the affected status (must be binary). The allele case-control test and genotypic case-control test, implemented using a contingency table analysis, are described in Nielsen and Weir (1999). The allele test assumes HWE. The test statistics have a chi-square distribution. Note that the degrees of freedom for the genotypic test will be the number of unique categories of genotypes examined in the data (either in case population or control population). In some cases, this number will not be the same as the theoretical number of genotypes ($=k(k+1)/2$, where k is number of alleles). The third method implemented in

the package, the multi-allelic trend test (Slager and Schaid 2001), has the same degrees of freedom as allele test but remains valid even with the violation of HWE assumption.

Simulation

SNP simulation

The coalescence simulation with the hotspot recombination model of Wiuf and Posada (2003) is implemented in PowerMarker. With no hotspot defined, the simulation becomes the classical coalescence model with homogenous recombination (Hudson 1983; Hudson and Kaplan 1990). Mutation is superimposed following an infinite-site model. The user can define a variety of parameters for the hotspot recombination model and the infinite-site model. The optional output of the module includes genealogy trees, the simulated probability distribution function of recombination rate, haplotypes and genotypes.

APPENDIX A: Estimating population specific F-statistics

This appendix uses the notation and concepts described by Weir and Hill (2002). Define an indicator variable x_{ijku} for the k th allele of the i th individual in the j th population

$$= \begin{cases} 1 & \text{if allele is type } A_u \\ 0 & \text{otherwise.} \end{cases}$$

Then, population specific F-statistics θ_i , between-population F-statistics $\theta_{ii'}$ and population specific total inbreeding coefficient F_i are defined as the correlation between x_{ijku} and $x_{i'j'k'u}$:

$$\varepsilon(x_{ijku}) = p_u$$

$$\varepsilon({x_{ijku}}^2) = p_u$$

$$\varepsilon(x_{ijku}, x_{i'j'k'u}) = \begin{cases} p_u^2 + p_u(1-p_u)\theta_{ii'} & i \neq i', j \neq j', k \neq k' \\ p_u^2 + p_u(1-p_u)\theta_i & i = i', j \neq j', k \neq k' \\ p_u^2 + p_u(1-p_u)F_i & i = i', j = j', k \neq k' \end{cases}.$$

Define

$$n_\cdot = \sum_{i=1}^{r} n_i$$

$$n_{ic} = n_i - \frac{\sum_{i=1}^{r} n_i^2}{r n_\cdot}$$

$$\tilde{p}_{iu} = \frac{1}{2n_i} \sum_{j=1}^{n_i} \sum_{k=1}^{2} x_{ijku}$$

$$\tilde{p}_u = \frac{1}{n_\cdot} \sum_{i=1}^{r} n_i \tilde{p}_{iu}$$

$$\pi_u = p_u(1-p_u)$$

$$\phi_i = \theta_i + \frac{1}{2n_i}(1 + F_i - 2\theta_i)$$

So that

$$\varepsilon(\tilde{p}_{iu}) = p_u$$
$$\varepsilon(\tilde{P}_{iuu}) = P_{iuu} = p_u^2 + \pi_u F_i$$
$$Var(\tilde{p}_{iu}) = \pi_u \theta_i + \frac{1}{2n_i}\pi_u(1 + F_i - 2\theta_i) = \pi_u \phi_i$$
$$Cov(\tilde{p}_{iu}, \tilde{p}_{i'u}) = \pi_u \theta_{ii'}$$
$$\varepsilon(\tilde{p}_u) = p_u$$
$$Var(\tilde{p}_u) = \frac{\pi_u}{n_.^2}\sum_{i=1}^{r}\phi_i n_i^2 + \frac{1}{n_.^2}\sum_{i=1}^{r}\sum_{i'\neq i}^{r} n_i n_{i'}\theta_{ii'}.$$

If the terms in the mean square for individuals within populations (MSI) and for alleles within individuals (MSG) are weighed by n_{ic} instead of n_i , then the sum of squares corresponding MSP , MSI and MSG have expectations

$$\varepsilon(SSP) = 2\varepsilon(\sum_{i=1}^{r} n_i(\tilde{p}_{iu} - \tilde{p}_u)^2) = 2\varepsilon(\sum_{i=1}^{r} n_i \tilde{p}_{iu}{}^2 - (\sum_{i=1}^{r} n_i)\tilde{p}_u^2)$$
$$= 2\pi_u[\sum_{i=1}^{r} n_{ic}\phi_i - \frac{1}{n_.}\sum_{\substack{i,i'=1 \\ i'\neq i}}^{r} n_i n_{i'}\theta_{ii'}]$$
$$\varepsilon(SSI) = \varepsilon(\sum_{i=1}^{r} n_{ic}(\tilde{p}_{iu} + \tilde{P}_{iuu} - 2\tilde{p}_{iu}^2)) = \pi_u[\sum_{i=1}^{r} n_{ic} - \sum_{i=1}^{r} n_{ic}(2\phi_i - F_i)]$$
$$\varepsilon(SSG) = \varepsilon(\sum_{i=1}^{r} n_{ic}(\tilde{p}_{iu} - \tilde{P}_{iuu})) = \pi_u[\sum_{i=1}^{r} n_{ic} - \sum_{i=1}^{r} n_{ic}F_i],$$

suggesting that π_u can be estimated as

$$\hat{\pi}_u = \frac{SS_u}{2(1-\theta_A)\sum_{i=1}^{r} n_{ic}}, \text{ where } SS_u = SSP + SSI + SSG, \theta_A = \frac{\sum_{\substack{i,i'=1 \\ i\neq i'}}^{r} n_i n_{i'}\theta_{ii'}}{\sum_{\substack{i,i'=1, \\ i\neq i'}}^{r} n_i n_{i'}}.$$

Therefore, from the expectations

$$\varepsilon(\sum_{u=1}^{m}\tilde{p}_{iu}(1-\tilde{p}_{iu}))=(\sum_{u=1}^{m}\pi_u)(1-\phi_i)$$

$$\varepsilon(\sum_{u=1}^{m}(\tilde{p}_{iu}-\tilde{P}_{iuu}))=(\sum_{u=1}^{m}\pi_u)(1-F_i)$$

$$\varepsilon(\sum_{u=1}^{m}[\tilde{p}_{iu}(1-\tilde{p}_{i'u})+\tilde{p}_{i'u}(1-\tilde{p}_{iu})]=2(\sum_{u=1}^{m}\pi_u)(1-\theta_{ii'}).$$

A moment estimate of ϕ_i for independent populations is

$$\hat{\phi}_i=1-\frac{2(1-\theta_A)(\sum_{i=1}^{r}n_{ic})\sum_{u=1}^{m}\tilde{p}_{iu}(1-\tilde{p}_{iu})}{\sum_{u=1}^{m}SS_u}.$$

A moment estimate of F_i for independent populations is

$$\hat{F}_i=1-\frac{2(1-\theta_A)(\sum_{i=1}^{r}n_{ic})\sum_{u=1}^{m}(\tilde{p}_{iu}-\tilde{P}_{iuu})}{\sum_{u=1}^{m}SS_u}.$$

An estimate of $\theta_{ii'}$ is given by

$$\hat{\theta}_{ii'}=1-\frac{(1-\theta_A)(\sum_{i=1}^{r}n_{ic})\sum_{u=1}^{m}[\tilde{p}_{iu}(1-\tilde{p}_{i'u})+\tilde{p}_{i'u}(1-\tilde{p}_{iu})]}{\sum_{u=1}^{m}SS_u}.$$

Define $\alpha_i=\frac{F_i-\theta_A}{1-\theta_A},T_i=\frac{\phi_i-\theta_A}{1-\theta_A},\beta_i=\frac{\theta_i-\theta_A}{1-\theta_A},\beta_{ii'}=\frac{\theta_{ii'}-\theta_A}{1-\theta_A}$, $f_i=\frac{F_i-\theta_i}{1-\theta_i}=\frac{\alpha_i-\beta_i}{1-\beta_i}$,

then θ_A is not involved in the estimates of $a_i,\beta_i,\beta_{ii'},$ $f_{i,}$ and T_i:

$$\hat{a}_i = 1 - \frac{2(\sum_{i=1}^{r} n_{ic}) \sum_{u=1}^{m} (\tilde{p}_{iu} - \tilde{P}_{iuu})}{\sum_{u=1}^{m} SS_u}$$

$$\hat{T}_i = 1 - \frac{2(\sum_{i=1}^{r} n_{ic}) \sum_{u=1}^{m} \tilde{p}_{iu}(1 - \tilde{p}_{iu})}{\sum_{u=1}^{m} SS_u}$$

$$\hat{\beta}_{ii'} = 1 - \frac{(\sum_{i=1}^{r} n_{ic}) \sum_{u=1}^{m} [\tilde{p}_{iu}(1 - \tilde{p}_{i'u}) + \tilde{p}_{i'u}(1 - \tilde{p}_{iu})]}{\sum_{u=1}^{m} SS_u}.$$

From the definition $\phi_i = \theta_i + \frac{1}{2n_i}(1 + F_i - 2\theta_i) \Rightarrow \beta_i = \frac{n_i}{n_i - 1} T_i - \frac{a_i}{2(n_i - 1)} - \frac{1}{2(n_i - 1)}$,

then $\hat{\beta}_i = 1 - \frac{2(\sum_{i=1}^{r} n_{ic}) \sum_{u=1}^{m} (\frac{n_i}{n_i - 1} \tilde{p}_{iu}(1 - \tilde{p}_{iu}) - \frac{1}{2(n_i - 1)} (\tilde{p}_{iu} - \tilde{P}_{iuu}))}{\sum_{u=1}^{m} SS_u}$.

These estimates can be simplified by defining

$$S_1 = \sum_{u=1}^{m} SS_u$$

$$S_{2i} = 2n_c \sum_{u=1}^{m} (\frac{n_i}{n_i - 1} \tilde{p}_{iu}(1 - \tilde{p}_{iu}) - \frac{1}{2(n_i - 1)} (\tilde{p}_{iu} - \tilde{P}_{iuu}))$$

$$S_{3i} = 2n_c \sum_{u=1}^{m} (\tilde{p}_{iu} - \tilde{P}_{iuu}),$$

then equations for estimating population specific F-statistics are

$$\hat{a}_i = 1 - \frac{S_{3i}}{S_1}$$

$$\hat{\beta}_i = 1 - \frac{S_{2i}}{S_1}$$

$$\hat{f}_i = \frac{S_{2i} - S_{3i}}{S_{2i}}.$$

The weighed average can be estimated as

$$\hat{a}_W = \frac{\sum_{i=1}^{r} n_i a_i}{n.} = 1 - \frac{\sum_{i=1}^{r} n_i S_{3i}}{(n.)S_1}$$

$$\hat{\beta}_W = \frac{\sum_{i=1}^{r} n_i \beta_i}{n.} = 1 - \frac{\sum_{i=1}^{r} n_i S_{2i}}{(n.)S_1}$$

$$\hat{f}_W = \frac{\sum_{i=1}^{r} n_i f_i}{n.} = \frac{\sum_{i=1}^{r} n_i S_{2i} - \sum_{i=1}^{r} n_i S_{3i}}{\sum_{i=1}^{r} n_i S_{2i}}.$$

For equal sample sizes these equations reduce to the estimators given by Weir and Cockerham (Weir and Cockerham 1984).

APPENDIX B: List of frequency-based distances

Let p_{ij} and q_{ij} be the frequencies of i th allele at the j th locus in populations X and Y respectively, while a_j is the number of alleles at the j th locus, and m is the number of loci examined.

Geometric distances are not negative, symmetric and satisfy the triangle inequality. The most common distance is the Euclidean distance, defined as:

$$D_{EU} = \frac{1}{m}\sum_{j}^{m}\sqrt{\sum_{i}^{a_j}(p_{ij} - q_{ij})^2}.$$

Rogers's (1972) distance is a scaled Euclidian distance:

$$D_R = \frac{1}{m}\sum_{j}^{m}\sqrt{\frac{1}{2}\sum_{i}^{a_j}(p_{ij} - q_{ij})^2}.$$

Prevosti *et al.*'s (1975) distance has statistical properties similar to those of D_R and is defined as:

$$C_p = \frac{1}{2m}\sum_{j}^{m}\sum_{i}^{a_j}| p_{ij} - q_{ij} |.$$

Cavalli-Sforza and Edwards' (1967) distance gives the chord distance between the two populations if we represent two populations on the surface of a multidimensional hypersphere using allele frequencies at the j th locus:

$$D_C = \frac{2}{\pi m}\sum_{j=1}^{m}\sqrt{2(1 - \sum_{i=1}^{a_j}\sqrt{p_{ij}q_{ij}})}.$$

Bhattacharyya (1946) and Nei (1987) recommended that the distance between the two populations be measured by

$$\theta^2 = \frac{1}{m}\sum_{j=1}^{m}(\arccos\sum_{i=1}^{a_j}\sqrt{p_{ij}q_{ij}})^2.$$

The Sanghvi distance (1953) was derived from chi-square goodness-of-fit statistics, and the distance is defined as:

$$X^2 = \frac{2}{m}\sum_{j=1}^{m}\sum_{i=1}^{a_j}\frac{(p_{ij}-q_{ij})^2}{(p_{ij}+q_{ij})}.$$

Nei *et al.*'s (1983) D_A distance:

$$D_A = 1-\frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{a_j}\sqrt{p_{ij}q_{ij}}\ .$$

None of the geometric distances described above involve any evolutionary models. Assuming that there is no mutation, and that all gene frequency changes are by genetic drift alone, the following two quanties are expected to rise linearly with amount of genetic drift.

Cavalli-Sforza's chord distance (1969) is given by:

$$f_v = 2\sqrt{\frac{\sum_{j=1}^{m}\left(1-\sum_{i=1}^{a_j}\sqrt{(p_{ij}-q_{ij})^2}\right)}{\sum_{j=1}^{m}(a_j-1)}}.$$

Reynolds, Weir, and Cockerham's (1983) genetic distance (ignoring the terms involving sample size n) is:

$$\theta_W = \frac{\sum_{j=1}^{m}\sum_{i=1}^{a_j}(p_{ij}-q_{ij})^2}{2\sum_{j=1}^{m}(1-\sum_{i=1}^{a_j}p_{ij}q_{ij})}.$$

Nei's (1972) standard distance has an expected value linearly related to the time since divergence, assuming that all loci have the same rate of neutral mutation, and that the genetic variation is maintained by the equilibrium between infinite-alleles mutation and genetic draft, with the effective population size of each population remaining constant. The quantity is defined as:

$$D_S = -\ln(J_{XY} \Big/ \sqrt{J_X J_Y}),$$

where $J_X = \sum_{j=1}^{m} \sum_{i=1}^{a_j} p_{ij}^2 / m$, $J_Y = \sum_{j=1}^{m} \sum_{i=1}^{a_j} q_{ij}^2 / m$, $J_{XY} = \sum_{j=1}^{m} \sum_{i=1}^{a_j} p_{ij} q_{ij} / m$.

Nei's (1973) minimum genetic distance (D_m), Latter's (1972) ϕ^* distance, and Latter's (1973) D_L distance are all defined similarly:

$$D_m = (J_X + J_Y)/2 - J_{XY}$$
$$\phi^* = \frac{(J_X + J_Y) - J_{XY}}{1 - J_{XY}}$$
$$D_L = -\ln(1 - \phi^*).$$

With the stepwise mutation model (SMM) assumption, Goldstein et al. (1995) proposed that the following distance be used for microsatellite loci:

$$(\delta\mu)^2 = \frac{1}{m} \sum_{j=1}^{m} (\mu_{X_j} - \mu_{Y_j})^2,$$

where $\mu_{X_j} (= \sum_k k p_{kj})$ and $\mu_{Y_j} (= \sum_k k q_{kj})$ are the average numbers of repeats found, and p_{kj} and q_{kj} are the frequencies of the allele with k repeats at the j th locus in population X and population Y, respectively.

A distance measure closely related to $(\delta\mu)^2$ is the average square distance (*ASD*, Slatkin 1995), which is given by

$$ASD = \frac{1}{m}\sum_{j=1}^{m}\sum_{u,v}(u-v)^2\, p_{uj}q_{vj}\ .$$

Another related distance measure is Shriver et al.'s (1995) distance, defined as

$$D_{SW} = W_{XY} - (W_X + W_Y)/2\ ,$$

where

$$W_X = \frac{1}{m}\sum_{j=1}^{m}\sum_{u,v}|u-v|\, p_{uj}p_{vj}\ , W_Y = \frac{1}{m}\sum_{j=1}^{m}\sum_{u,v}|u-v|\, q_{uj}q_{vj}\ , W_{XY} = \frac{1}{m}\sum_{j=1}^{m}\sum_{u,v}|u-v|\, p_{uj}q_{vj}$$

Another commonly used distance, the shared allele distance D_{SA} (Chakraborty and Jin, 1993), is defined as:

$$D_{SA} = \frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{a_j}\min(p_{ij}, q_{ij})\ .$$

The measure $D_{LS} = -\ln(1-D_{SA})$ (usually referred as log shared allele distance) has also been proposed.

APPENDIX C: Haplotype estimation

The EM algorithm (Excoffer and Slaktkin 1995) is an iterative method to reconstruct haplotypes and find frequencies (F) that maximize the likelihood of the genotype data. Under the assumption of Hardy-Weinberg equilibrium, the likelihood is the product of the probabilities of each individual conditional on haplotype frequencies:

$$L(F) = \prod_{i=1}^{n} \Pr(G_i \mid F),$$

where G_i is the genotype of the i th individual and n is the sample size.

Define S_i as the set of ordered pairs of haplotypes that constitute the genotype G_i . The E-step refreshes the genotype frequencies from the haplotype frequencies as follows:

$$P_i = \sum_{[j,j']\in S_i} p_j p_{j'},$$

where $[j, j']$ is the ordered pair of j th haplotype and j' th haplotype, and p_j and $p_{j'}$ are the current frequencies of j th and j' th haplotypes, respectively..

At the M-step, maximal likelihood estimates of these haplotype frequencies are obtained and used in turn as the haplotype frequencies at the next iteration:

$$p_k = \frac{1}{2n} \sum_{i=1}^{n} \sum_{[j,j']\in S_i} \frac{m_{jj'} p_j p_{j'}}{P_i},$$

where

$$m_{jj'} = \begin{cases} 2 & \text{if } j = j' = k \\ 1 & \text{if } j \neq j', j = k \text{ or } j' = k. \\ 0 & \text{otherwise} \end{cases}$$

The E- and M-step are iterated until the likelihood $L(F)$ converges.

The following optimizations are used in my implementation of the EM algorithm:

(1) If all loci are biallelic and the number of loci (m) is smaller than 32, a haplotype is stored as an integer. Otherwise it is stored as an array of integers.

(2) If $10 < m < 32$ or one of the locus is not biallelic, the whole haplotype pool (H) is stored in a hashtable, with the key and the value storing the haplotype and the haplotype information (e.g frequency of the haplotype), respectively. If all loci are biallelic and $m \le 10$, the information for the haplotype h_r is directly stored at the r th position of the assigned storage (2^m positions), where r is the integer value of the haplotype. The direct storage mechanism will decrease the complexity of common operations (such as search and insertion) to constant time. For each unique genotype, S_i is stored as a collection of pairs of pointers to the haplotype pool. Therefore, for each genotype the E-Step and M-Step can be performed by enumerating S_i instead of by enumerating the whole (larger) haplotype pool.

(3) If $m \ge 32$, a "key" tree is used to store the haplotype pool. A haplotype can be constructed by connecting all its branches from the root of the tree to the terminal. The tree structure improves the performance in two aspects. First, a haplotype tree is a compressed storage for the haplotype pool. More importantly, a tree structure allows "branch-and-bound" algorithms to be performed.

Statistical methods for unrelated population data consider all possible haplotype pairs consistent with the independent genotypes and provide the complete haplotype pool, which might be larger than the real one. Xiang *et al.* (2003) proposed a modified EM algorithm to infer the haplotype frequencies using parent-offspring trios and showed the

new algorithm is superior to the usual EM algorithm that uses only independent parents or children. However, their method still considers the complete haplotype pool. In PowerMarker we developed a new approach to efficiently estimate the MLEs of the haplotype frequencies for the children. Incorporating pedigree information can help solve some ambiguous phases and eliminate a majority of impossible haplotypes, which improves both the accuracy and the capability of statistical methods. Consider the trio design, where all children are assumed to be independent. The constraint of parental information on the possible phase of the single child will significantly reduce the number of compatible haplotype pairs for each child. The EM algorithm is then applied to the independent children's genotype data, based on the reduced haplotype set S_i^C for i th individual:

$$\text{E-Step: } P_i = \sum_{[j,j']\in S_i^C} p_j p_{j'},$$

$$\text{M-Step: } p_k = \frac{1}{2n}\sum_{i=1}^{n}\sum_{[j,j']\in S_i^C}\frac{m_{jj'}p_j p_{j'}}{P_i}.$$

To estimate the haplotype frequency of the parents, we have to take recombination into account. A reasonable assumption is that for the chromosomal region investigated, at most one recombination event occurs per meiosis. An optimal procedure to do the parental haplotype estimation, as well as the diagnosis of the extended EM algorithms, is currently under investigation.

A**CKNOWLEDGEMENTS**

This program has been made possible by NSF grants DBI-0096033 and DEB-9996118. Many thanks to: Ed Buckler, Xiang Yu, Li Li, John Doebley, Yves Vigouroux, Kenji Fukunaga, and all other users or beta-testers of PowerMarker that have sent us their comments.

REFERENCES

Botstein D, White RL, Skolnick M and Davis RW, Construction of a genetic linkage map in man using restriction fragment length polymorphisms, *American Journal of Human Genetics*, **32**: 314-331 (1980).

Bhattacharyya A, On a measure of divergence between two multinomial populations. Sankhya 7: 401-407 (1946).

Cavalli-Sforza LL, Human diversity. Proc. 12th Intl Cong. Genet., Tokyo 3: 405-416 (1969).

Cavalli-Sforza LL and Edwards AWF, Phylogenetic analysis: models and estimation procedures, *American Journal of Human Genetics*, **19**: 233-257 (1967).

Devlin B and Risch N, A comparison of linkage disequilibrium measures for fine-scale mapping, *Genomics*, **29**: 311-322 (1995).

Excoffier L and Slatkin M, Maximum-Liklihood estimation of molecular haplotype frequencies in a diploid population, *Molecular Biology and Evolution*, **12**: 921-927 (1995).

Felsenstein J, Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39**: 783-791.

Felsenstein J, PHYLIP (phylogeny Inference Package), version 3.5c. Depart of Genetics, University of Washington, Seattle.

Goldstein DB and Ruiz Linares A, Cavalli-Sforza LL and Feldman MM, Genetic absolute dating based on microsatellites and the origin of modern humans. *Proc. Natl. Acad. Sci.* USA **92**: 6723-6727 (1995).

Guo SW and Thompson EA, Performing the exact test of Hardy-Weinberg proportion for multiple alleles, *Biometrics*, **48**: 361-372 (1992).

Hudson RR, Properties of a neutral allele model with intragenic recombination. *Theor. Pop. Biol.* **23**: 183-201 (1983).

Hudson RR and Kaplan N, Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**: 147-164 (1990).

Latter BDH, Selection in finite populations with multiple alleles. III. Genetic divergence with centripetal selection and mutation. *Genetics*, **70**: 475-490 (1972).

Nei M, Genetic distance between populations, *American Naturalist*, **106**: 283-292 (1972).

Nei M, The theory and estimation of genetic distance, p45-54 in Genetic Structure of Populations, edited by Morton NE, University Press of Hawaii, Honolulu (1973).

Nei M, Molecular Evolutionary Genetics. Columbia University press, New York (1987).

Nei M and Takezaki N, Estimation of genetic distances and phylogenetic trees from DNA anlysis. Proc. 5th World Cong. Genet. Appl. Livstock Prod. 21: 405-412 (1983).

Nielsen DM and Weir BS, A classical setting for associations between markers and loci affecting quantitative traits, *Genetic Research*, **74**: 271-277 (1999).

Prevosti A, Ocana J and Alonzo G, Distances between populations for *Drosophila Subobscura* based on chromosome arrangement frequencies. *Theo. Appl. Genet.* **45**: 231-241 (1975).

Reynolds J, Weir BS and Cockerham CC, Estimation of the Coancestry coefficient: basic for a short-term genetic distance. *Genetics* **105**: 767-779 (1983).

Rogers JS, Measures of genetic similarity and genetic distance, pp. 145-153 in *Studies in Genetics* VII. University of Texas Publication 7213, Austin, TX (1972).

Sanghvi LD, Comparison of genetical and morphological methods for a study of biological differences. *Amer. J. Phys. Anthropol.* **11**: 385-404 (1953).

Shriver M, Jin L, Boerwinkle E, Ferrell R et al., A novel measure of genetic distance for highly polymorphic tandem repeat loci. *Mol. Biol. Evol.* **12**: 914-920 (1995).

Slager SL and Schaid DJ, Evaluation of candidate genes in case-control studies: A statistical method to account for related subjects. *American Journal of Human Genetics*, **68**: 1457-1462 (2001).

Slatkin M, A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139**: 457-462 (1995).

Weir BS, Genetic data analysis II, Sunderland, MA: Sinauer Associates, Inc (1996).

Weir BS and Hill WG, Estimating F-Statistics, *Annu. Rev. Genet.* **36**: 721-750 (2002).

Wiuf Carsten and Posada David, A coalescent model of recombination hotspots. To appear in *Genetics* (2003)

Yu X, Nielson DM and Weir BS, An EM Algorithm for Haplotype Frequency Estimation Using Parents-offspring Trios. Submitted to *American Journal of Human Genetics* (2003).

Zaykin D, Zhivotovsky L and Weir BS, Exact tests for association between alleles at arbitrary numbers of loci, *Genetica*, **96**: 169-178 (1995).