# ASSESSING GENETIC DIVERSITY IN GERMPLASM COLLECTIONS USING MOLECULAR MARKERS

by

R. W. Doerge
Purdue University

B. W. S. Sobral
National Center for
Genome Resources

Technical Report #97-13

Department of Statistics
Purdue University
West Lafayette, IN USA

August, 1997

# Assessing Genetic Diversity in Germplasm Collections Using Molecular Markers

R.W. Doerge[1] and B.W.S. Sobral[2]

August 1997

[1] Departments of Agronomy and Statistics, 1399 Mathematical Science Building
Purdue University, West Lafayette, IN 47907
[2] National Center for Genome Resources, 1800 Old Pecos Trail, Santa Fe, NM 87505

**Running Head:** Estimation of Core Collections from Germplasm Collections

**Corresponding author:** R.W. Doerge
Department of Statistics
1399 Mathematical Sciences Building
Purdue University
West Lafayette, IN 47907-1399

**Telephone number:** (765) 494-6030
**Fax number:** (765) 494-0558
**E-mail:** doerge@stat.purdue.edu

**Key words:** germplasm collection, biodiversity
core collection, molecular markers
conversation biology

## Abstract

Genetic and cultural diversity have shaped the state of current *ex situ* germplasm collections, not only in the way genetic material is collected, but also in the manner genetic resources are utilized. The size and maintenance of germplasm collections are issues of great concern among geneticists and plant breeders. Reduction in size of current germplasm collections, allied with an increase in use of the new smaller germplasm collections has long-reaching financial and functional impacts. If characterized, smaller, yet representative collections provide functionality by supplying new alleles for breeding programs that are directed toward specific agronomic traits. Genotypic breeding strategies using genomic information rely heavily on molecular genetic data as a means of advancement, by identifying genomic regions associated with agronomic traits, as well as in comparative mapping of traits and in genomes across related species.

The task undertaken by this research effort relies on computer simulation of DNA marker data to represent a large germplasm collection under varying conditions, for the purpose of reducing the simulated collection to a smaller collection (known as a "core collection") while retaining genetic diversity. The reduction mechanisms studied rely on two sampling techniques, both of which are based upon a specified number of DNA markers, and allowable similarity between individuals. The first sampling procedure is regulated by choosing individuals which maximize genetic dissimilarity. This fits the rather unrealistic scenario of genotyping all individuals (at a large number of loci, using dominant markers) in a entire collection and then choosing the most diverse set of individuals. The second sampling procedure chooses randomly from the total collection, and represents the scenario of picking random individuals to represent the diversity of the entire collection. Genetic diversity is measured in terms of the variance of pairwise Jaccard similarity indices. The long-term hope of this simulation study is to forecast the impact that recent advances in molecular genetic technology can have on diversity assessment, size reduction, maintenance, development, and use of world germplasm collections.

# Introduction:

Efforts toward conservation and retention of the world's genetic resources are drawing together some of the top researchers in the scientific community (Bataillon *et al.* 1996, Virk *et al.* 1996, Clark *et al.* 1997). With the world's supply of natural resources being challenged by increasing population size and decreasing reserve from which to regain our genetic base, proactive human intervention is of paramount importance. For hundreds of years genetic resources have been collected in what are now referred to as germplasm collections. These resources initially were chosen based upon many significant characteristics which reflected the need and direction of the existing society. Over time, these collections have grown increasingly large and counterproductive for use by breeders. Frankel (1984) suggested the creation of core collections that would represent the genetic diversity in the species and its relatives while possessing a minimum number of repetitive members. Brown (1995) gave the following formal definition for a *core collection*

> "A *core collection* consists of a limited set of accessions (sample maintained in the whole collection) derived from an existing germplasm collection, chosen to represent the genetic spectrum in the whole collection. The core should include as much as possible of its genetic diversity."

The actual steps toward creating core collections has generated a great amount of discussion (Hodgkin *et al.* 1995), yet the implementation of these steps remain to be seen in total.

The short–term goal is to assess existing germplasm collections to reduce duplication and to extract a subset (core) that can be intensively studied, characterized, and utilized. The core can then be allowed regulated growth through additional genetic contribution, rather than repetitive genetic resources. Essentially, the core collection is an active collection, while the larger collection is maintained mainly for storage purposes. The long–term goal is to retain the majority of the overall genetic variation, not the majority of individuals. These points are both the motivation and reason for this study. If the first step in maintaining genetic diversity is evaluating what exists in current collections, then it needs to be done in an efficient and cost–effective manner.

For many years man has selected for superior characteristics of individuals within a small range of species represented by major crops and some of their relatives. Initially, these selection decisions were based on historical understanding of the crop, and experience gained through continued breeding. As a results, many breeders, and many years of breeding experience and decision have shaped the structure of existing germplasm collections. Today's germplasm collections are large and represent cultural decisions based upon societies that may no longer exist, as well as breeding rules that have benefited from years of refinement. In addition, germplasm collections are not supported by any structural model, or even a set of rules invoked by breeders that explains the formation of the collection as it evolved. Based upon the lack of structural architecture that created the germplasm collections, attempting to statistically model the underlying genetic basis represented in the total of the collection is practically impossible. As a result, there is no defined structure in current collections that allows breeders to efficiently explore sources of new alleles for traits of interest. Conservation and management of existing collections is the key to maintaining current levels of diversity. Selection based upon cultural (DNA) decisions may now be coupled with the extraordinary advances that have been made in molecular genetic technology.

DNA marker technology has provided a powerful vehicle for the genetic characterization of germplasm collections around the world. In the context of this research, molecular markers are used to assess genetic diversity of large germplasm banks, and have potential to be viewed as a mechanism to reduce collection size. The task seems simple enough: using a large number of molecular markers, identify each individual in the collection by genotyping, and retain only those individuals that provide the maximum amount of genetic diversity. By doing this, a smaller "core" collection may be retained as a representation of the whole. These "core" collections could then, themselves, be described through the amount of genetic variation they contain. The decision to add an individual to the core collection could be evaluated based upon its genetic similarity to the collection, not added solely upon physical or geographical features. In the context of predicting performance, molecular markers coupled with quantitative trait measurements may be assessed for associations to predict the performance of a sample of germplasm. Virk *et al.* (1996) used associations between quantitative traits and RAPD markers to predict performance traits for samples of *Oryza sativa* (Asian rice) germplasm. Approaching these two issues within the DNA marker framework brings to bare many important questions. How many molecular markers are required? How does the frequency of the marker in the population affect the size of the core collection? What measure of genetic diversity should be used in conjunction with the DNA data? Can we afford to do this? Is genotyping the entire collection necessary, or is there a more efficient way to meet our goal? While germplasm collection diversity assessment, size control and predicting trait performance are equally important issues, we will concentrate on the first, and leave the latter to future research.

Currently, there is a dichotomy pertaining to how one should evaluate and maintain information kept on the individuals in existing gene banks or germplasm collections. Historically, breeders evaluate numerous aspects pertaining to each individual maintained in a collection. With this information summarized, each member of the collection has a corresponding "passport" of relevant information based solely on physical features or collection features. Passport data ranges in quality and quantity. This data may include only the collection site, with little additional information, or it may include a rich description of morphological characteristics, as well as indigenous names and uses. Unfortunately, because of movement of plant genotypes by humans, passport data may suggest two different genotypes when only one exists, with different names or, conversely, it may suggest the same genotype, based on the same name, when, in fact, the materials are genetically distinct. Thus, passport data is not a reliable indicator as these data are frequently incomplete or of dubious quality. In addition, passport information rarely allow discrimination of genetic similarity, except as measured through morphology. Molecular geneticists, on the other hand, maintain that DNA marker technology should be exploited for the purpose of identifying individuals genetically, allowing a higher resolution than morphology alone. The method of choosing the "best" or most representative individuals in the collection remains to be decided. Using this as our motivation, our goal becomes genetic diversity characterization of existing collections. We are not suggesting the elimination of passport data in any way; we suggest the inclusion of genotypic data in the passport. What is being put forth is a method which will genetically characterize individuals for which trait or other characteristics may exist via passport information. In the end, the more characterization, molecular as well as agronomic and botanical, the better.

The purpose of this paper is to address assessment of genetic diversity in existing germplasm collections. Our motivation for this work was the following statement by Virk *et al.* (1996)

> "The management of (such) collections is difficult simply because of their vast size, and there is a clear requirement for the development of procedures which utilize fast and

> reliable methods for the measurement of diversity in order to facilitate the organization and prioritization of germplasm resources."

While division of whole collections into subcollections using passport information has been suggested (Mackay 1995), our purpose is to assess genetic diversity in a statistically sound manner. We address this situation from a data simulation standpoint where the amount of variation/genetic diversity represented in the current germplasm collection is known through simulated molecular marker data.

## Materials and Methods:

Biochemical and DNA markers may be used to characterize genetic diversity in germplasm collections. Each of the many molecular and biochemical marker technologies encompass positive and negative aspects, of which cost, speed, and simplicity are of vital (Gepts 1995) importance when one considers the size of the task at hand. For these reasons, we will initiate this study of genetic diversity (Virk *et al.* 1995) using randomly amplified polymorphic DNA (RAPD) (also known as arbitrarily primed PCR markers) (Welsh and McClelland 1990, Williams *et al.* 1990). We realize that other DNA markers provide more information, and we plan to extend this investigation to such cases in the near future. Although various types of DNA marker systems exist, each with its advantages and limitations, the general situation for most plant genetic resources, with the exception of those that have strong agroindustrial interests, is one of limiting funds. This is especially true for crops that are important to the developing world, many of which are called "orphan crops". Many of these crops are vegetatively reproduced, meaning that the expenses associated with maintenance of the collection are exacerbated. Together with the limited funds for their study, orphan crops pose the most serious problem for creation, maintenance and use of *ex situ* germplasm collections. Yet, because of their relatively undomesticated state, they are the ones that stand to gain the most from methodologies that enhance the use of germplasm to create new varieties. Among the tropical orphan crops, many are polyploid. In polyploids, the main limitation of arbitrary primer methods, namely the dominant phenotype associated with the marker system, is unimportant because current methodologies for genetic mapping and QTL identification are based on necessary use of dominant marker systems, whatever their nature. In the case of polyploidy, this is an imposition of the genetic nature, not an inherent characteristic of the marker system *per se.*

Recent studies (Schoen and Brown 1993; Holsinger 1993; Milligan *et al.* 1994) have shown use of molecular makers to determine single locus variation through a variety of sampling schemes. Both allozyme and DNA markers have been used to retain high levels of neutral alleles in a subsample of a base collection. Use of neutral alleles to dissect genetic variation has been called into question by Bataillon *et al.* (1996). The most recent attempt to study the sampling mechanism for germplasm collection reduction was put forth by Bataillon *et al.* (1996) using computer simulation to compare the efficacy of various sampling regimes for the purpose of retaining neutral and nonneutral alleles in a host of different environments, along with numerous conditions on mutation, genetic drift, and gamete formation. Our purpose in this work is not to repeat the already informative work of Bataillon *et al.* (1996), but rather present a different viewpoint on the sampling mechanism used to assess genetic similarity, variation, diversity, etc. based on a simple DNA marker system, and a limited number of restrictions.

**Similarity Index:** Although it is well known that various conditions and forces have directed the evolution of the species represented in the world's many germplasm collection, any attempt to model these forces is just that, an attempt. In some cases models of drift, mutation, migration and selection have projected intriguing results and hypotheses that have made great differences in our understanding of the evolutionary process. However, in this work we wish to maintain a minimum number of forces (conditions) which explain the genetic variation in an attempt to simulate, what some may consider a more realistic view of germplasm collections, while some may consider it largely naive and uninformed. Since there is no genetic structure (*e.g.*, experimental design) imposed on germplasm collections, attempting to represent any genetic similarity through modeled molecular marker information is not considered. This point was recently stated by Bataillon *et al.* (1996):

> "It is seldom if ever possible to assess in a comprehensive manner the amount and structure of genetic variation in a population or collection of interest to genetic conservation, making it difficult to proceed in a rational way toward construction of representative samples for conservation."

With this previous point as our direction, we choose a reliable similarity measure which imposes no restrictions on the amount and structure of genetic variation. We realize, of course, that any one of many similarity measures could be used.

For any single individual in the collection, $m$ markers may be genotyped in a dominant manner, such that the information per individual is a series of presences or absences (binary responses) of each of the $m$ markers. Let 1 indicate the presence of a marker with frequency $p$, and 0 with frequency $(1-p)$, otherwise. The comparison of two individuals, $i$ and $j$ across all markers $(m_{ij} = a_{ij} + b_{ij} + c_{ij} + d_{ij})$ provides the following $2 \times 2$ table

| | i= 1 | i=0 |
|---|---|---|
| j=1 | a$_{ij}$ | b$_{ij}$ |
| j=0 | c$_{ij}$ | d$_{ij}$ |

Individuals $i$ and $j$ have the same marker state (1) at $a_{ij}$ of the $m$ markers with probability $p^2$, while they differ at $b_{ij} + c_{ij}$ of the $m - a_{ij}$ remaining markers with probability $2p(1-p)$. The value $d_{ij}$ represents all reasons two markers fail to show signal (missing value or a true null). Treating the $0-0$ matches as irrelevant, we use Jaccard's similarity (1908) index as our measure. Let

$$J_{ij} = \frac{a_{ij}}{a_{ij} + b_{ij} + c_{ij}} \tag{1}$$

represent the similarity between individuals $i$ and $j$, across $m$ markers each of which occur at some frequency $0.0 \leq p_l \leq 0.50$; $l = 1, \cdots, m$. In the long run, with fixed marker frequency, we expect the Jaccard index to have value:

$$E[J] = \left[\frac{m}{m-(a+b+c)}\right] \frac{p^2}{p^2 + 2p(1-p)},$$

and variation

$$Var[J] = \left[\frac{m}{m-(a+b+c)}\right] \frac{2p(1-p)}{(2-p)^2},$$

where $m \neq a + b + c$.

In its extreme values, Jaccard's domain is

$$0.0 \leq J \leq 1.0,$$

and may be interpreted as follows. If two individuals share no matches in their markers, $J = 0$, and they are considered unique. Perfect correspondence between individuals ($J = 1$) is achieved when there are no mismatches ($b + c = 0$) and no irrelevant information ($d = 0$). As the measure, $J$, increases, the similarity between individuals (across markers) increases. Based upon this assessment, Jaccard's measure between individuals serves our overall purpose of characterizing germplasm collections using a dominant marker system.

We chose to use summary information Jaccard's similarity indices in our assessment of the overall genetic diversity between individuals in a large germplasm collection. In this regard, some investigators consider heterozygosity as a more accurate evaluation, since it considers allele frequency across multiple markers of varying allelic states. Milligan *et al.* (1994) show that heterozygosity estimates can be made, but more markers must be sampled along with some adjustments. Since we implement a binary (dominant) marker system heterozygosity is masked. Given that we have a large number of individuals, and a potentially large pool of genetic markers, Jaccard's measure supports the use of arbitrarily primed PCR (RAPD) technologies as fast, cost efficient techniques to characterize our germplasm collection. Since there is no genetic structure imposed on germplasm collections, attempting to represent any modeled genetic similarity is unrealistic. No genetic assumptions are attached to Jaccard's measure, thus there are no parametric implications.

Assuming there are $N$ individuals in a germplasm collection, $N!/2$ possible pairwise similarity indices are required to assess the overall similarity between each member of the collection, and when taken together, summarize the genetic diversity of the collection. The expected overall Jaccard measure in a collection represents the average amount of similarity between all pairwise comparisons of individuals in a particular collection. The variation among the $N!/2$ Jaccard measures represents the dispersion of similarity among the pairwise individual comparisons. Based upon the variance of $N!/2$ Jaccard measures one can identify, for the collection, a maximum amount of "dissimilarity" between individuals or diversity. Conceptually, the variation in the Jaccard measure may be used as an unique measure of diversity in a germplasm collection. Together with a specific sampling scheme, it may be possible to use the variation in the pairwise Jaccard measures to meet our goal of representing the most diversely associated individuals in a smaller core collection without actually genotyping every member of the collection at a large number of markers. An initial assumption that we will discuss again later in the paper is that genetic diversity *does* exist in the large germplasm collection. That is to say, not all individuals are identical, nor are all the individuals similar in the same way (*e.g.*, all the same, or all very unique). For all intended purposes, this assumption is not outside the realistic parameters of existing collections.

## Sampling Schemes:

Two sampling strategies are adopted for the purpose of assessing genetic diversity of germplasm collections. A **maximum diversity** sample is obtained for comparison with a **random** sample in order to demonstrate relative efficiency of one over the other. We explain both sampling mechanisms in turn, the means of comparison, and then we described the actual germplasm simulations.

**Maximum Diversity Sample (MDS)**: Realistically, current germplasm collections maintain a number of repetitive individuals or at least extremely similar individuals. Theoretically, the entire collection can be characterized molecularly for all individuals, across a large number of markers, and comparisons made between individuals. Individuals providing the largest amount of diversity (unique alleles) can be retained in a core collection, with the repetitive resources either released or maintained in a less expensive manner (*i.e.*, long–term storage). For example, it is estimated that roughly 30% of the Cassava (*Manihot esculenta* Craz.) national collection in Brazil is reiterated. The Cassava collection is maintained largely in the field, at great expense and labor. If true, a simple identification of duplicates could immediately release resources for studies aimed at utilization of genetic resources in Cassava breeding. It is well known that regeneration from storage (seed) is cheaper than even one year of clonal storage. While this sampling scheme is easy to understand, and even justify, our argument is that it is not realistic, and other other avenues should be pursued in the interest of utility, time, and money. In order to regulate the size of the maximum dissimilar sample (*i.e.*, individuals chosen from the total collection as being diverse), so that it is less than the total collection, the sample size is restricted by the level of genetic similarity specified, say $J \leq 2\%$. In detail, the entire collection is genotyped at $m$ markers, and all pairwise Jaccards are calculated between all individuals. The decision to create a core collection of $n$ individuals who represent up to a certain amount of similarity between them is made. If the pairwise Jaccards for the entire population are ordered (retaining each individual's identity) by magnitude, then each individual that meets the Jaccard requirement with at least one other individual is sampled. In the final maximum diversity sample, every individual sampled is represented only once, and the sample can not be larger that the initial collection. Since the sample is determined by the limit placed on the Jaccard measure, sample size is determined intrinsically, and is an important issue in the comparison between MDS and random sampling.

**Random Sample (RS):** Based upon the fundamentals of Statistics, a random sample, if large enough will represent the total of a larger, unobtainable population. This concept is critical to statistical inferences concerning population parameters. The extension to germplasm characterization is straightforward since the amount of genetic diversity in a germplasm collection is a parameter describing the collection. Rather than perform costly molecular procedures on the total collection, or growing out all individuals side-by-side for evaluation, both very time-consuming and costly procedures, random sampling (RS) may provide an alternative. The intention of this investigation is to explore RS. A random sample, of the same sample size determined by MDS, is taken from the whole of the germplasm collection and genotyped at the equivalent number of markers that would have been performed on the entire collection. Issues concerning sample size will be discussed later in this paper.

**Comparison of MDS and RS:** Whether the mechanism to assess diversity is the total characterization of the entire collection, and then choose a top percentage of diverse individuals (MDS), or a random sample (RS), the Jaccard measure is used to summarize the (dis)similarity between individuals. The variation in this measure can be computed under both schemes (MDS and RS), and can be used as a measure of diversity in the total simulated collection, the maximum diverse sample, and the random sample. For example, if all individuals in the collection are extremely similar (the same holds for extremely dissimilar individuals), then the pairwise Jaccard measures will be large (close to 1.0) or small (close to 0.0), respectively, and the variation among these pairwise Jaccard measures small. On the other hand, if the total collection represents a large amount of diversity among the individuals, then the pairwise Jaccard measures may encompass the full range of Jaccard values, and the variation in the Jaccard measures from all pairwise comparisons may be large. Employing our previously stated assumptions on the available diversity in the collection, we use the variation of the Jaccard measure to assess the diversity of the collection, not to assess the amount of similarity. This means that a collection with a large number of similar individuals will provide the same low variation among the Jaccard measures as a large number of dissimilar individuals, and will be represented as such through the variation of the Jaccard measure. For comparative purposes the sample size representing the most diverse individuals in the collection and a random sample must be equivalent, as explained previously. The measure of variation under both schemes (equivalent sample size) will serve as evidence of worth in actual implementation of either strategy. Our anticipation is that random sampling will serve relatively well as an alternative means for assessing genetic diversity of the collection, the question of how well remains.

### Germplasm Simulation:

The use of simulated germplasm data (collection) is essential in this framework. Simulation allows the liberty of knowing everything (*e.g.*, marker frequency, marker number, sample size, etc.) about the collection, and thus allows us the ability to assess how well the collection is described through MDS and RS.

A large germplasm collection with 5000 individuals is simulated under two scenarios, increasing uniform marker frequency ($0 \leq p \leq 0.5$ presented; $0.50 < p \leq 1.0$ not presented), and random marker frequency (same range and presentation) across increasing marker number. Binary marker states represent arbitrarily primed PCR markers (*e.g.*, RAPDs). The average and variance of the Jaccard measure in the entire germplasm collection is calculated from pairwise comparisons between 5000!/2 individuals. Samples of the total collection are taken according to the described schemes, with MDS regulating the sample size as the allowable similarity ranges from 0 to 10%. Marker number increases from 40 to 240 in increments of 40. Additional markers are evaluated per increment of marker number, after the full range of allowable similarity (for that marker number) is completed. Initially, 5000 individuals and 40 markers are simulated, MDS implemented under a specified amount of similarity to determine sample size, and RS implemented at the same sample size. The variation under each scenario of the respective sampling schemes is calculated, and compared. The variance of the pairwise Jaccards for the entire simulated population is also calculated for the purpose of assessing how well MDS and RS perform in representing the true simulated scenario. The simulation process continues with the addition of 40 more markers being evaluated in the total germplasm collection, and the sampling schemes used to create respective core collections. The addition of increasing markers in the simulation process is a realistic approach to the larger problem at hand. Generally, more markers are added to the evaluation process until a majority of the genome is covered.

## Results:

The results of the described simulations are presented for random marker frequencies and fixed marker frequencies under increasing marker number (40, 80, 120, 160, 200) and Jaccard limit (0,.10). Maximum diversity sampling and random sampling are described and compared within each section.

### Random Marker Frequency:

**Sampling Distribution of the Jaccard Measure:** We consider the sampling distribution of the Jaccard measure for random marker frequencies using a small simulation study for the purpose of investigating the effect of marker frequency on the Jaccard measure across increasing marker number. 50 random individuals were simulated for 40, 80, 120, 160, and 200 markers occurring with random frequency, 1225 (50 choose 2) pairwise Jaccard measures calculated and plotted to form the histograms shown in Figure 1. No notable change, in the center of distribution of the Jaccard measure as the marker number increases, is observed. As expected, as the marker number increases, the sampling distribution demonstrates less variation.

**Sample Size:** As the sample size increases (as determined by MDS), the amount of similarity (using Jaccards) between individuals increases for a fixed marker number (40, 80, 120, 160, 200), see Figure 2. As the marker number increases, smaller samples represent lower levels of similarity between the individuals. In situations where we are required to take an extremely conservative sample size, these simulations demonstrate the limitation on the amount of similarity under increasing marker number. However, for increasing marker number, one can easily observe (Figure 2) that smaller sample sizes adequately assess the amount of similarity between individuals. For example, from Figure 2, one is able to consider a random sample of less than 500 individuals (from a collection of 5000) scored at 160 dominant markers which allows up to 3% similarity in the core collection. Meaning that the 500 individuals in the collections are different at 97% of the markers scored.

**Random Sampling:** For comparative purposes (with MDS, and the total collection), the variance of the Jaccard measures in the random sample is plotted against the Jaccard limit that determined the samples size (in MDS). It would have been equally informative to plot variation against sample size, since sample size is determined by allowable levels of similarity. As the similarity increases, $(0.0 \leq J \leq 0.10)$, the amount of diversity (variance of the Jaccard measures) in the random samples, over increasing marker number simulated with random frequencies (Figure 3a-3d), approaches the diversity of the total germplasm collection (because the sample size is increasing to meet the total population, see Figures 2). In addition, the amount of variability assessed in the core collection decreases as the number of markers increases (Figure 3b). It should not go without notice (Figure 3d) the point at which the number of markers increases, and the variability in the random sample under and over estimates the population variation (*i.e.*, m=160 and m=200). However, the amount of variation in the total of the germplasm collection is extremely small, for random marker frequencies, so the demonstration of over and under estimation of the variability is not of high relevance for this simulation, or this work.

**Maximum Diversity Sampling:** The behavior of maximum dissimilar sampling is the same as RS (Figure 3c), however the estimation of the total germplasm collection diversity is under estimated up to the point where marker number is 160 and 200. Again, in the instances where the estimated diversity begins to fluctuate, the total germplasm variability is so small that, like RS, the point is irrelevant. Figures 3a-3d demonstrate these last points well.

**Comparison of RS and MDS:** When compared to the total amount of diversity available in the whole germplasm collection, RS and MDS each equally represent the variance of the Jaccard pairwise measure, and thus the genetic diversity, regardless of marker number or similarity in the smaller collection. Recall, MDS requires that all of the germplasm collection be genotyped at numerous markers before the top percentage of the collection is chosen, and that the process is expensive and highly labor intensive. Relative to MDS, RS essentially does equally well in assessing the genetic diversity of the total collection through the variance of the Jaccard measures, but does so in a less expensive and labor intensive manner.

## Fixed Marker Frequency:

**Sampling Distribution of the Jaccard Measure:** Figures 4-8 show the sampling distributions for Jaccard using 1225 pairwise Jaccard measures simulated under the increasing marker number (same as RS) and fixed (and increasing) marker frequencies. As the marker frequency increases the center of the Jaccard distribution shifts toward 0.50. Conceptually, this makes sense, since the chance of observing the marker increases, then under increasing marker number, more information among individuals is available for assessment.

**Sample Size:** As the marker frequency increases for fixed marker number (Figure 9), the sample size approaches the total of the collection as the amount of similarity allowed between individuals increases. In other words, in order to sample the rarer genotypes, a larger sample has to be taken for markers of lower occurrence. Increasing marker number does not help if the marker frequency is low, which is not the case in most cultivated germplasm collections (e.g., the marker is so rare, that chances of collecting it at all is very small). However, as we approach more realistic conditions (marker frequency), the addition of more markers becomes increasing useful as a discriminatory tool. For example, when the marker frequency is 0.2 across all markers, we need to increase the marker number in order to make informed decision of realistic sample size (for a given Jaccard). For markers of higher frequency the chance of representing them in the sample is higher just by chance, however when the marker frequencies increase there is more power with more markers. The gain in benefit of more markers is extremely large, thus validating the usefulness of molecular technology.

**Random Sampling:** When the frequency of the markers increases from 0.0 to 0.5, over increasing marker number and increasing similarity, the available diversity in the random sample decreases as the similarity between individuals in the total collection increases (Figures 10b-13b). In fact, when the frequencies of the markers (regardless of number) is 0.50, no random sample can be drawn to represent any diversity in the total collections, within the pairwise (Jaccard) similarity measure range (0.0, 0.10). This is not to say that if the Jaccard range is extended, that a random sample at the 0.50 marker frequency could not be taken. Conceptually, this makes sense, since the marker genotypes are so frequent, all individuals have the same genotype, and therefore, no diversity exists in the collection for all markers of high occurrence.

**Maximum Diversity Sampling:** As the frequency of the markers increases from 0.0 to 0.5, over increasing marker number and increasing allowable similarity the same scenario that was just demonstrated with RS is seen with MDS (Figures 10c-13c). The ability to represent the diversity of the large collection becomes less as the occurrence of the markers become common place.

**Comparison of RS and MDS:** The comparison of sampling randomly from a simulated germplasm collection for genetic diversity to the sampling strategy which picks the most diverse individuals from the collections demonstrates little advantage to the latter scheme. When the average variation, across all Jaccard levels, is calculated for each marker number and frequency (Figure 10d-13d), and compared for the true simulated collection, the random sample, and the maximum diversity sample, RS and MDS do equally well in assessing the diversity in the simulated collection. However, the key objective of the work was to show the utility of one sampling scheme over the other with respect to cost and efficacy. To this end, RS requires less work and less money, solely because the entirety of the germplasm collection does not have to be genotyped in order to determine which individuals to maintain in the smaller core collection.

## Discussion:

The simulations presented here are limited in nature, and present only a focused view of the comparison between genotyping an entire germplasm collection and choosing the most diverse individuals, and taking a random sample of individuals. In an extended simulation that allowed marker frequencies to increase to 1.0, and the limit on the Jaccard condition to increase to 1.0, the general trends represented by the initial simulations continue. We chose to present the limited simulation in this manner because it demonstrated the utility of the sampling schemes in a population that has little variability. Many extensions to this simulation are possible, and include the use of codominant genetic markers, as well as initial simulated populations retaining a larger amount of variation across assessed marker genotypes. Our main hope in presenting this simulation study is not to make conclusive remarks based upon this one simulation, but rather to stimulate discussion, interaction, and progress in an area that deserves more attention.

## Efficient Use and Maintenance of Smaller Germplasm Collections:

Once the larger problem of assessing diversity and reducing germplasm collection size is addressed, the maintenance and use of the smaller collections must be regulated for the purpose of moving forward. The association between genomic regions (QTL, quantitative trait loci) and quantitative traits has received considerable attention over the last 15 years. The impact of smaller collections has enormous potential for the future identification of QTL. Once a core collection is well characterized with molecular markers and character data, regions of genomic association may be identified, and the total of this information maintained in a computer data base. Breeders may then determine a small number of markers for the trait they are interested in, return to the whole population, genotype those few markers across their population and select individuals from their population based upon what is already established in the core collection. Predicting the performance of any individual in the whole population based upon marker information on the core collection has long–ranging effects on the addition of material to core collection. When a potential new accession is found the genetic markers provide a way to genotype the accession and compare to the whole population, via the core collection, for the purpose of determining whether that individual accession is already in the population. Advancing in this manner allows the genotyping of the whole population in an unrestricted fashion, since the breeder has concentrated on specific markers. Under the alternative strategy of genotyping the entire population across a vast number of markers, there is a good chance that the markers the breeder needs for characterizing the trait of interest are not even identified in the population (*i.e.*, a finite number of markers will be used to represent the population). Ultimately, using a core collection and a random sampling strategy, more markers will be genotyped in the population (and linked to the trait) than if the whole population were initially genotyped for markers that have no potential use in the breeding. The future potential of this paradigm toward world germplasm collections is an interconnected data base linked to specific agronomically important traits. Supported by recent advances in comparative mapping, interconnected data bases linked across evolutionarily related species have huge potential for identifying genetic markers and genomic regions associated with with quantitative traits, thus simultaneously moving forward our understanding of genetic variation, regulation, and determination of many populations and species.

## Acknowledgments:

RWD and BWS thank Centro Internacional de la Papa (CIP) for providing the setting in which this work was cultivated. RWD thanks Drs. Bruce Craig, Marc Ghisilain, and Jaya Satagopan for helpful discussions during the initial stages of this work. BWS was supported in part by a grant from Copersucar (Brazil), and the Mauritian Sugar Industry (Re'duit).

## Cited Literature:

Bataillon, T.M., J.L. David, D.J. Schoen. 1996. Neutral genetic markers and conservation genetics: simulated germplasm collections. Genetics. **144**:409-417.

Brown, A.H.D. 1995. The core collection at the crossroads. In Hodgkin, T., A.H.D. Brown, Th.J.L van Hintum and E.A.V. Morales (eds) *Core Collections of Plant Genetic Resources.* West Sussex, UK: Wiley-Sayce.

Clark, R.L., H.L. Shands, P.K. Bretting, and S.A. Eberhart. 1997. Germplasm regeneration: developments in population genetics and their implications. Crop Science. **37**:1-6.

Frankel, O.H. 1984. Genetic perspectives of germplasm conservation. In Arber, W.K., Llimensee, K., Peacock, W.J. and Starlinger, P. (eds) *Genetic Manipulation: Impact on Man and Society.* Cambridge, UK: Cambridge University Press.

Gepts, P. 1995. Genetic markers and core collections. In Hodgkin, T., A.H.D. Brown, Th.J.L van Hintum and E.A.V. Morales (eds) *Core Collections of Plant Genetic Resources.* West Sussex, UK: Wiley-Sayce.

Holsinger, K.E. and R.K. Jansen. 1993. Phylogenetic analysis of restriction site data. In: *Molecular Evolution: Producing the Biochemical Data. Methods in Enzymology*, Vol 224, Zimmer E.A., White T.J., Cann R.L., Wilson A.C. (eds). New York: Academic Press.

Jaccard, P. 1908. Nouvelles recherches sur la distribution florale. Bull. Soc. Vaud. Sci. Nat., 44:223-270.

Mackay, M.C. 1995. One core collection or many? In Hodgkin, T., A.H.D. Brown, Th.J.L van Hintum and E.A.V. Morales (eds) *Core Collections of Plant Genetic Resources.* West Sussex, UK: Wiley-Sayce.

Milligan, B.G., J. Leebens-Mack J., A.E. Strand. 1994. Conservation genetics: beyond the maintenance of marker diversity. Mol Ecol **3**:423-435.

Hodgkin, T., A.H.D. Brown, Th.J.L. Hintum and E.A.V. Morales. (editors). 1995. Core Collections of Plant Genetic Resources. Wiley.

Schoen, D.J. and A.H.D Brown. 1993. Conservation of allelic richness in wild crop relatives is aided by assessment of genetic markers. Proc. Natl. Acad. Sci. USA **90**:10623-10627.

Virk, P.S., B.V. Ford-Lloyd, M.T. Jackson, H.S. Pooni, T.P. Clemeno, H.J. Newbury. 1996. Predicting quantitative variation within rice germplasm using molecular markers. Heredity. **76**:296-304.

Virk, P.S., B.V. Ford-Lloyd, M.T. Jackson, H.J. Newbury. 1995. Use of RAPD for the study of diversity within pant germplasm collections. Heredity. **74**:170-179.

Welsh, J. and M. McClelland. 1990. Fingerprinting genomes using PCR with arbitrary primers. Nucl. Acids Res., **18**:7213-7218.

Williams, J.G.K., A.R. Kubelik, K.J. Livak, J.A. Rafalski, and S.V. Tingey. 1990. DNA polymorphisms amplified by arbitrary primers use useful as genetic markers. Nucl. Acids Res. **18**:6531-35.

## Figure Headings:

**Figure 1:** Sampling distribution of pairwise Jaccard calculated for 50 individuals. Marker frequencies are random (0.0, 0.50), marker number 40, 80, 120, 160, 200 for figures 1a, 1b, 1c, 1d, 1e, respectively.

**Figure 2:** Effect of random marker frequency (0.0, 0.50), marker number, and specified limit of Jaccard measure on sample size. As the Jaccard limit increases to allow more similarity in the sample, larger samples are drawn.

**Figure 3:** Random marker frequency (0.0, 0.50), and Jaccard limit between (0.0, .10). Figure 3a represents the (Jaccard) variation in the total simulated population (*i.e.*, the truth) across increasing marker number. Figure 3b represents the (Jaccard) variation in the random sample (RS) of individuals drawn from the total simulated population. Figure 3c represents the maximum diverse sample (MDS) taken from the total population. Figure 3d is the average Jaccard variation across all marker numbers for the total population, the random sample, and the maximum diverse sample.

**Figure 4:** Sampling distribution of pairwise Jaccard calculated for 50 individuals. Fixed marker frequency p=.1, for increasing marker number.

**Figure 5:** Sampling distribution of pairwise Jaccard calculated for 50 individuals. Fixed marker frequency p=.2, for increasing marker number.

**Figure 6:** Sampling distribution of pairwise Jaccard calculated for 50 individuals. Fixed marker frequency p=.3, for increasing marker number.

**Figure 7:** Sampling distribution of pairwise Jaccard calculated for 50 individuals. Fixed marker frequency p=.4, for increasing marker number.

**Figure 8:** Sampling distribution of pairwise Jaccard calculated for 50 individuals. Fixed marker

frequency p=.5, for increasing marker number.

**Figure 9:** Effect of fixed marker frequency (a-d), marker number, and specified limit of Jaccard measure on sample size. As the frequency of the markers increases, the point at which the Jaccard limit converges for all marker numbers requires increasing sample sizes and a larger allowable Jaccard limit.

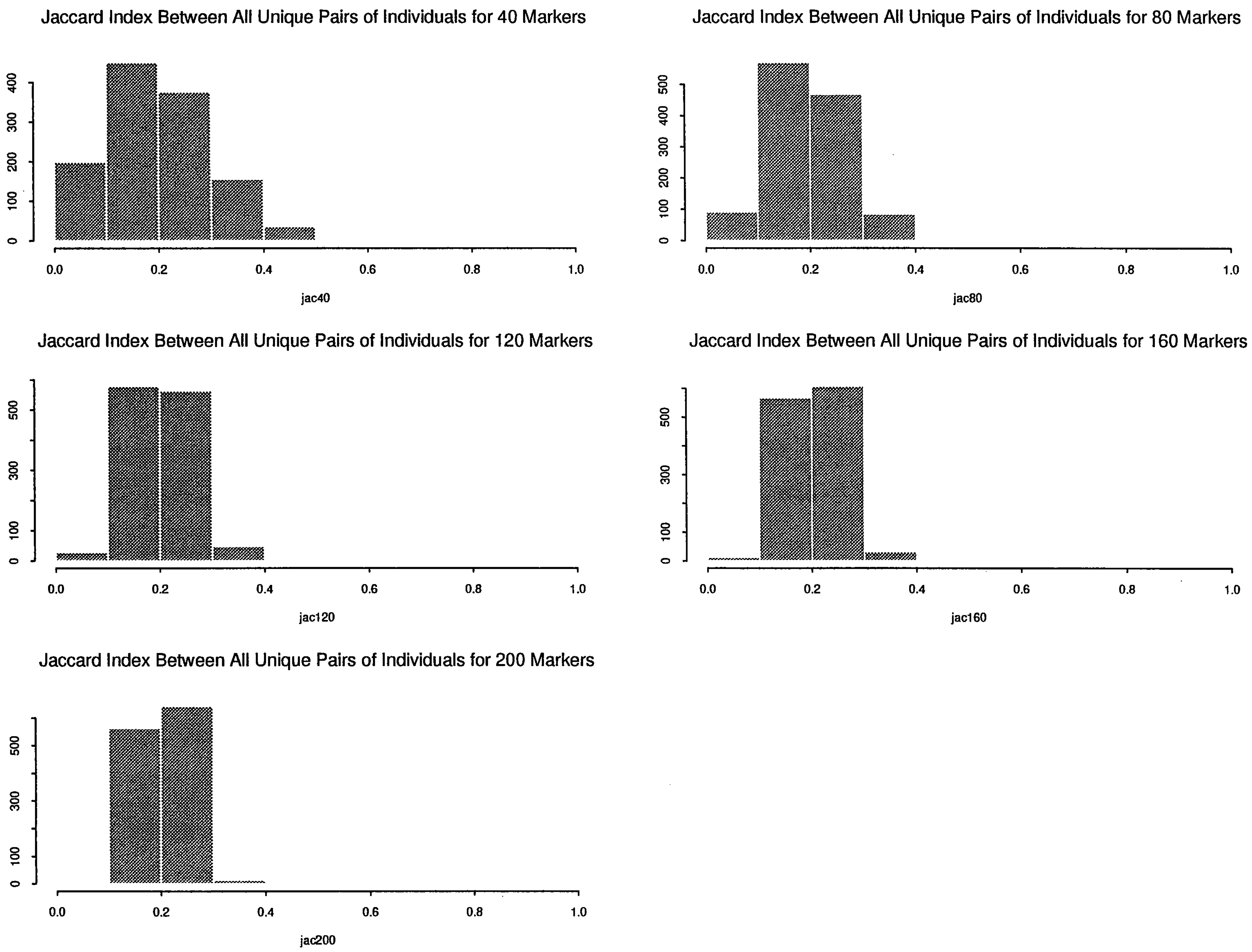
**Figure 10:** Fixed marker frequency p=.1, and Jaccard limit between (0.0, .10). Figure 10a represents the (Jaccard) variation in the total simulated population (*i.e.*, the truth) across increasing marker number. Figure 10b represents the (Jaccard) variation in the random sample (RS) of individuals drawn from the total simulated population. Figure 10c represents the maximum diverse sample (MDS) taken from the total population. Figure 10d is the average Jaccard variation across all marker numbers for the total population, the random sample, and the maximum diverse sample.

**Figure 11:** Fixed marker frequency p=.2, and Jaccard limit between (0.0, .10). Figure 11a represents the (Jaccard) variation in the total simulated population (*i.e.*, the truth) across increasing marker number. Figure 11b represents the (Jaccard) variation in the random sample (RS) of individuals drawn from the total simulated population. Figure 11c represents the maximum diverse sample (MDS) taken from the total population. Figure 11d is the average Jaccard variation across all marker numbers for the total population, the random sample, and the maximum diverse sample.

**Figure 12:** Fixed marker frequency p=.3, and Jaccard limit between (0.0, .10). Figure 12a represents the (Jaccard) variation in the total simulated population (*i.e.*, the truth) across increasing marker number. Figure 12b represents the (Jaccard) variation in the random sample (RS) of individuals drawn from the total simulated population. Figure 12c represents the maximum diverse sample (MDS) taken from the total population. Figure 12d is the average Jaccard variation across all marker numbers for the total population, the random sample, and the maximum diverse sample.

**Figure 13:** Fixed marker frequency p=.4, and Jaccard limit between (0.0, .10). Figure 13a

represents the (Jaccard) variation in the total simulated population (*i.e.*, the truth) across increasing marker number. Figure 13b represents the (Jaccard) variation in the random sample (RS) of individuals drawn from the total simulated population. Figure 13c represents the maximum diverse sample (MDS) taken from the total population. Figure 13d is the average Jaccard variation across all marker numbers for the total population, the random sample, and the maximum diverse sample.



Jaccard Index Between All Unique Pairs of Individuals for 40 Markers

Jaccard Index Between All Unique Pairs of Individuals for 80 Markers

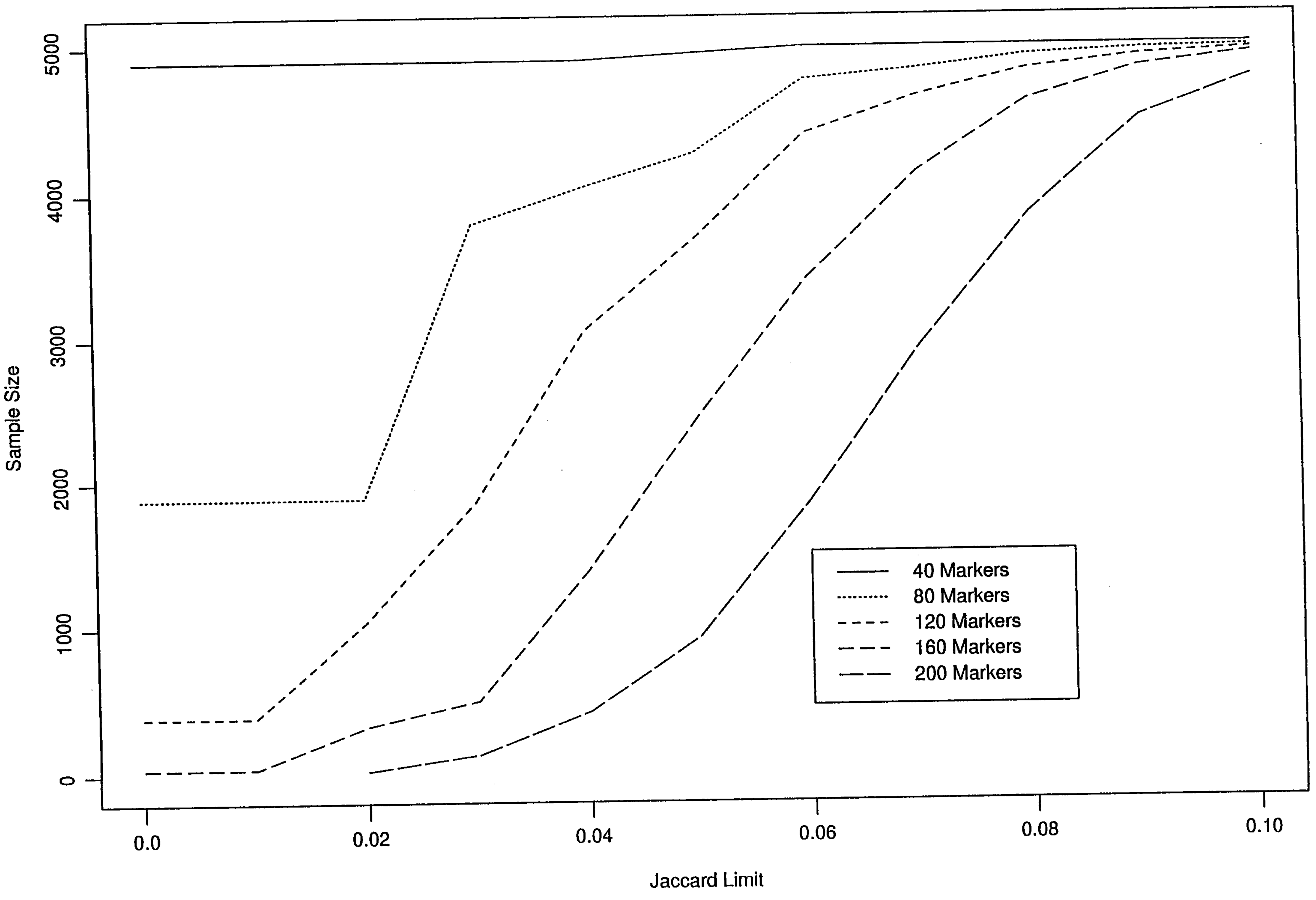
Jaccard Index Between All Unique Pairs of Individuals for 120 Markers

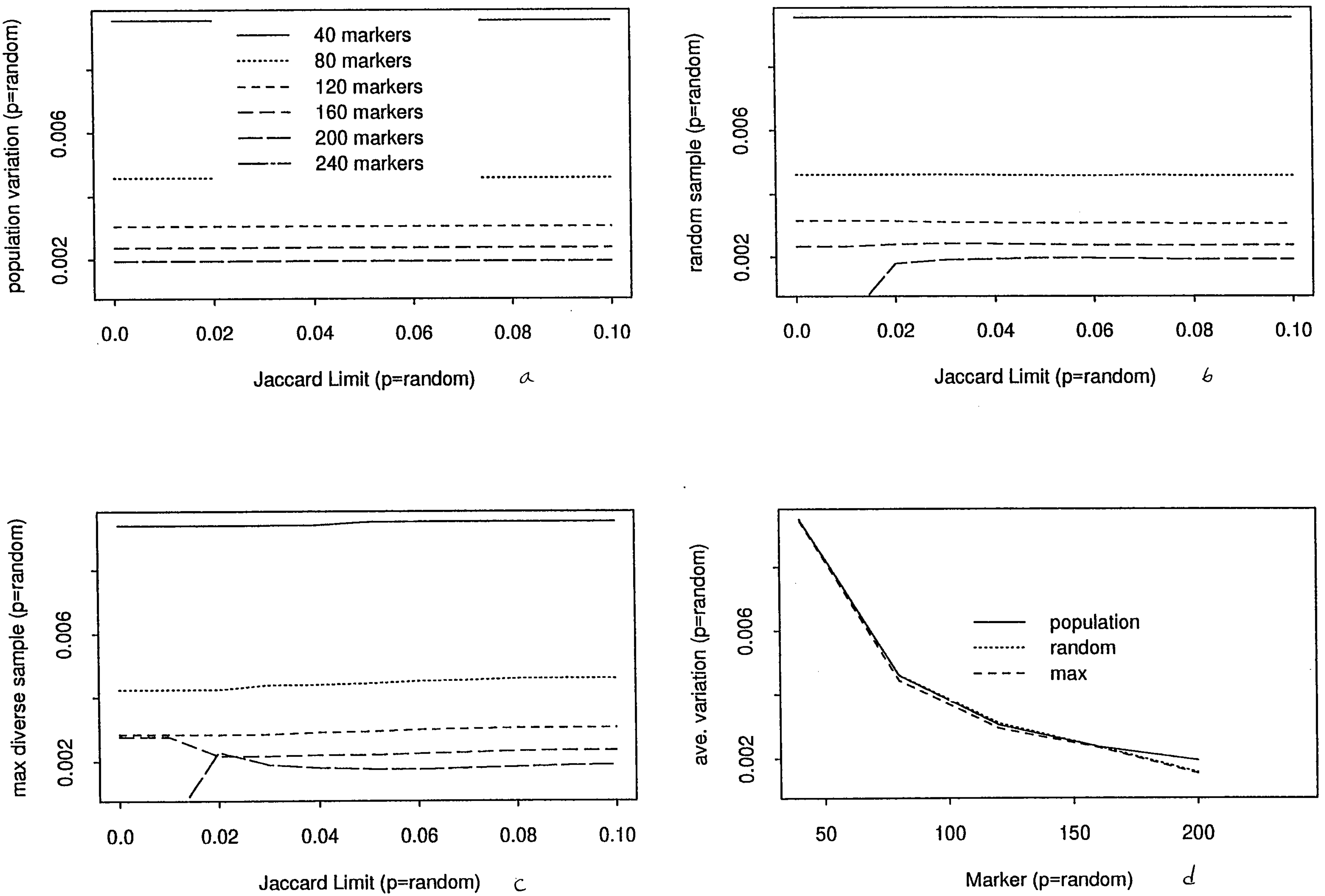
Jaccard Index Between All Unique Pairs of Individuals for 160 Markers

Jaccard Index Between All Unique Pairs of Individuals for 200 Markers

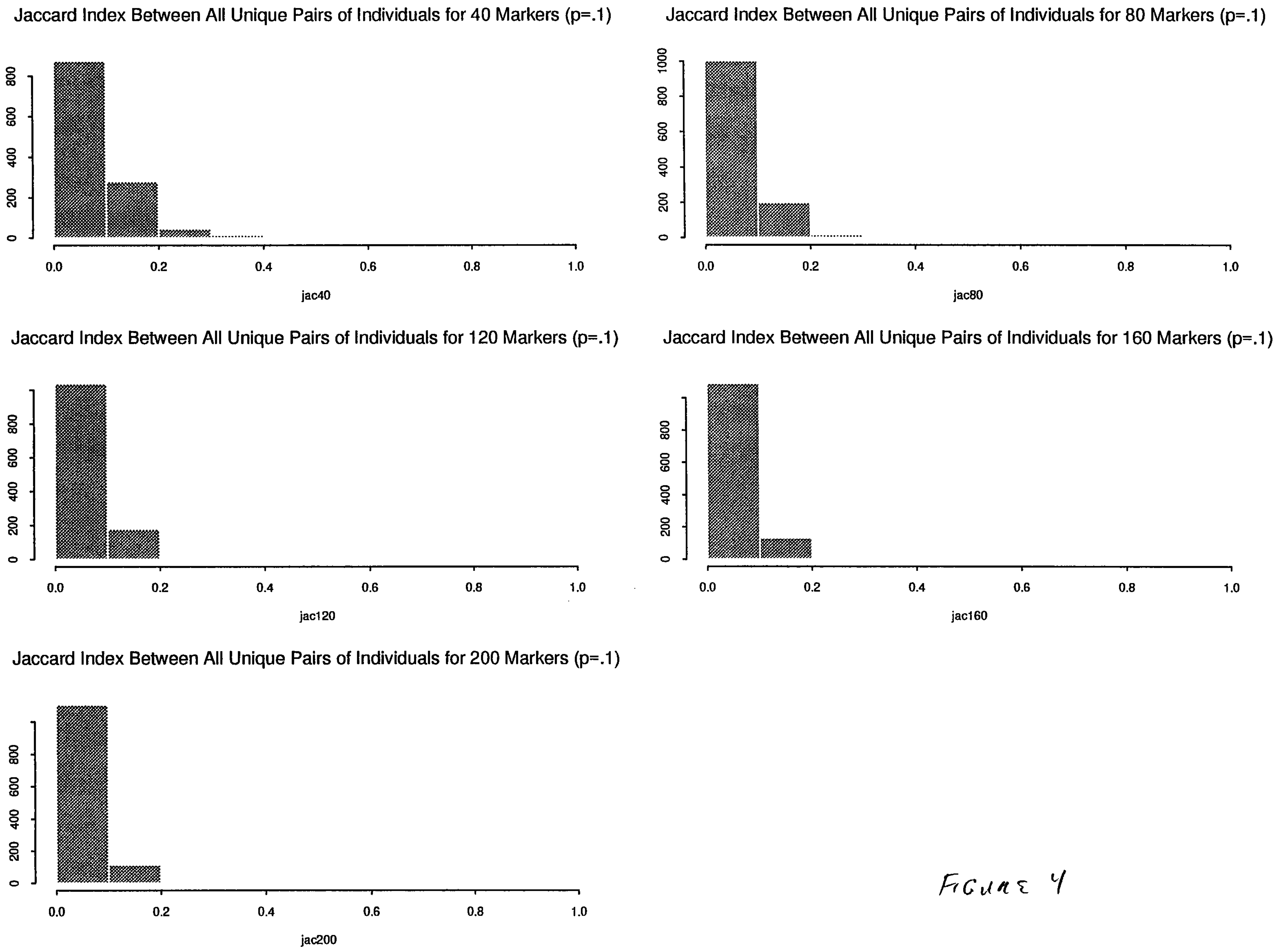
Figure 1

## Random Marker Frequency





population variation (p=random)
0.002
0.006
0.0
0.02
0.04
0.06
0.08
0.10
Jaccard Limit (p=random)
a
40 markers
80 markers
120 markers
160 markers
200 markers
240 markers
random sample (p=random)
0.002
0.006
0.0
0.02
0.04
0.06
0.08
0.10
Jaccard Limit (p=random)
b
max diverse sample (p=random)
0.002
0.006
0.0
0.02
0.04
0.06
0.08
0.10
Jaccard Limit (p=random)
c
ave. variation (p=random)
0.002
0.006
50
100
150
200
Marker (p=random)
d
population
random
max



Jaccard Index Between All Unique Pairs of Individuals for 40 Markers (p=.1)

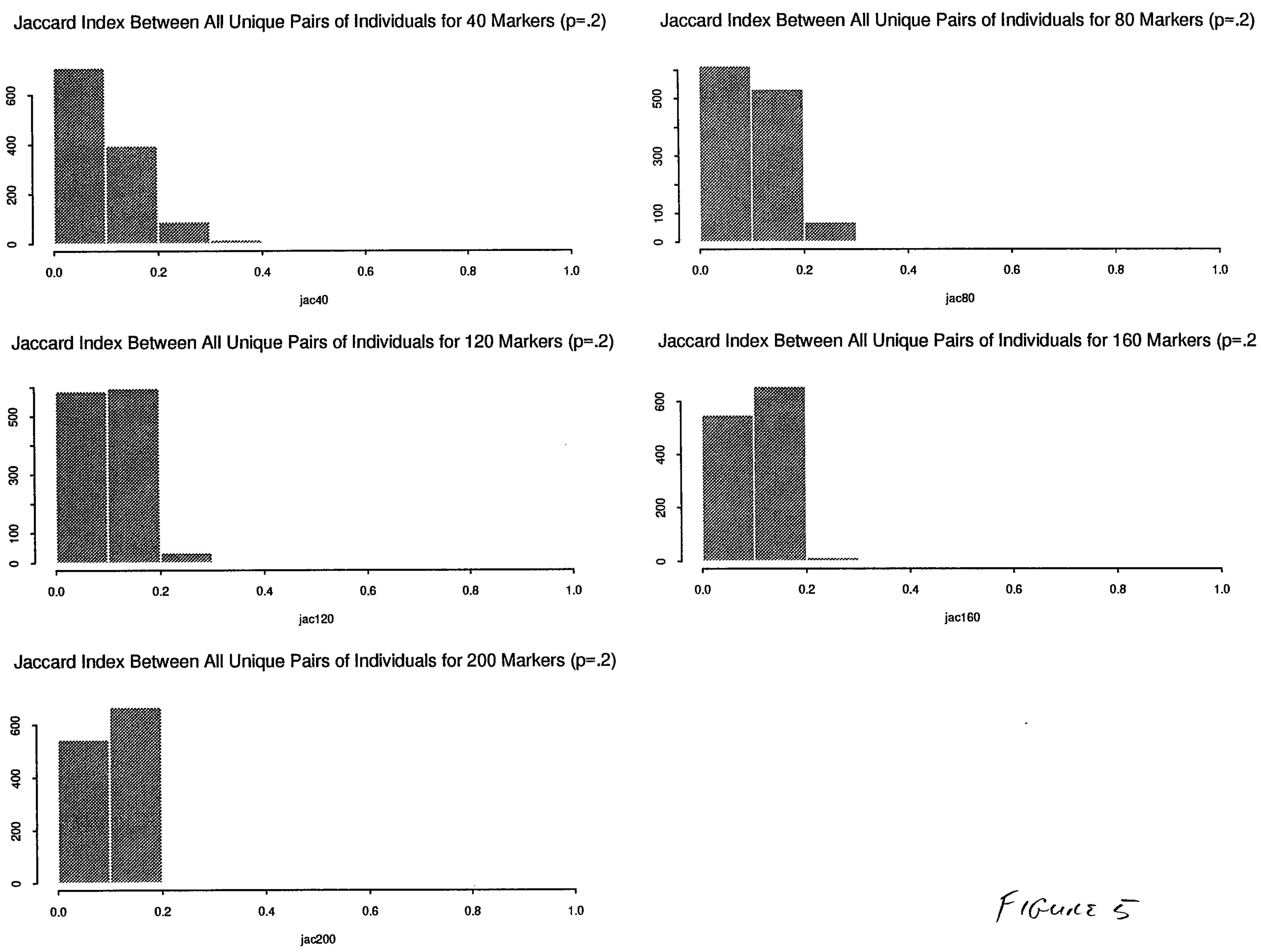
Jaccard Index Between All Unique Pairs of Individuals for 80 Markers (p=.1)

Jaccard Index Between All Unique Pairs of Individuals for 120 Markers (p=.1)

Jaccard Index Between All Unique Pairs of Individuals for 160 Markers (p=.1)

Jaccard Index Between All Unique Pairs of Individuals for 200 Markers (p=.1)

FIGURE 4



Jaccard Index Between All Unique Pairs of Individuals for 40 Markers (p=.2)

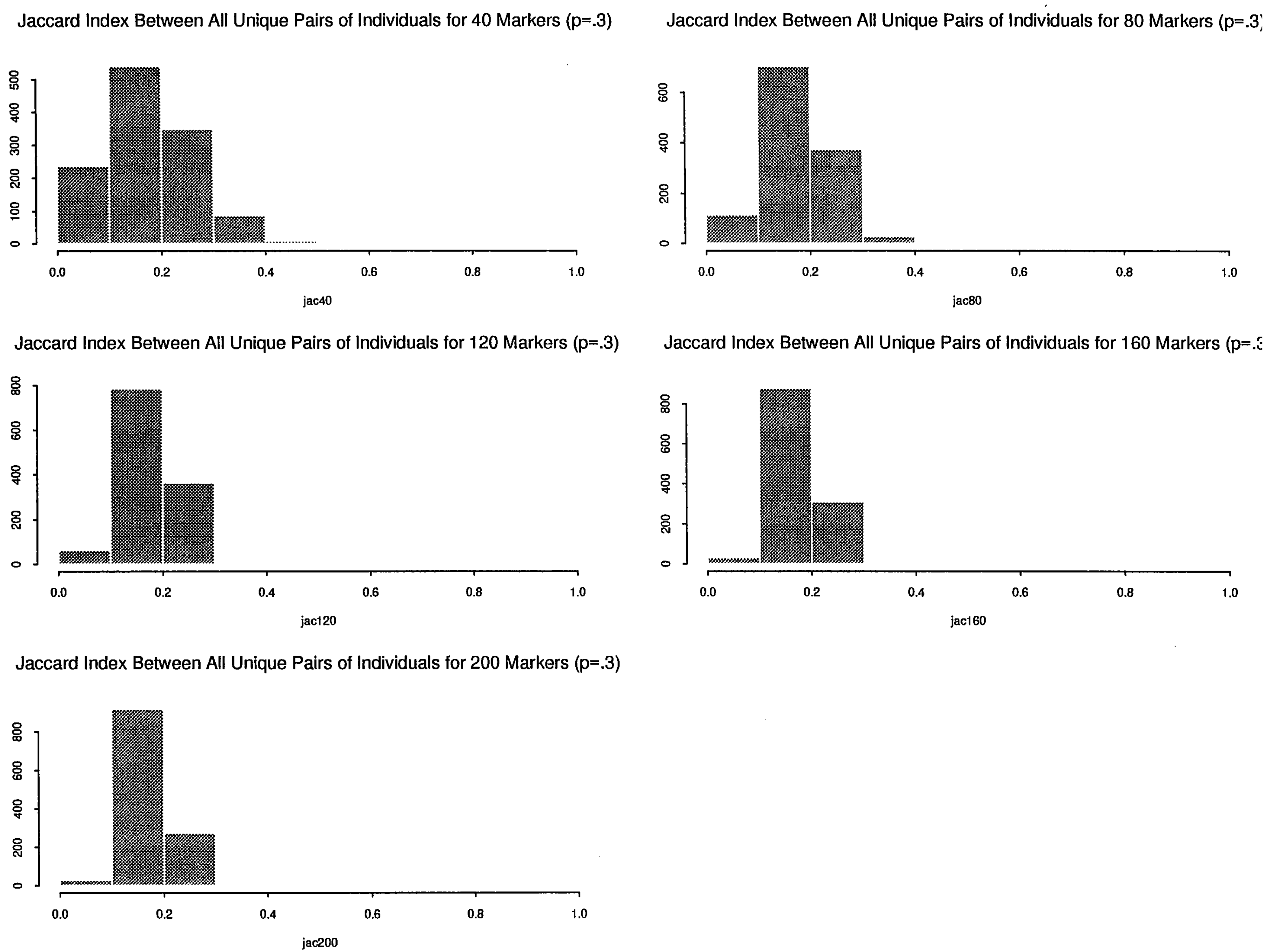
Jaccard Index Between All Unique Pairs of Individuals for 80 Markers (p=.2)

Jaccard Index Between All Unique Pairs of Individuals for 120 Markers (p=.2)

Jaccard Index Between All Unique Pairs of Individuals for 160 Markers (p=.2

Jaccard Index Between All Unique Pairs of Individuals for 200 Markers (p=.2)

FIGURE 5



Jaccard Index Between All Unique Pairs of Individuals for 40 Markers (p=.3)

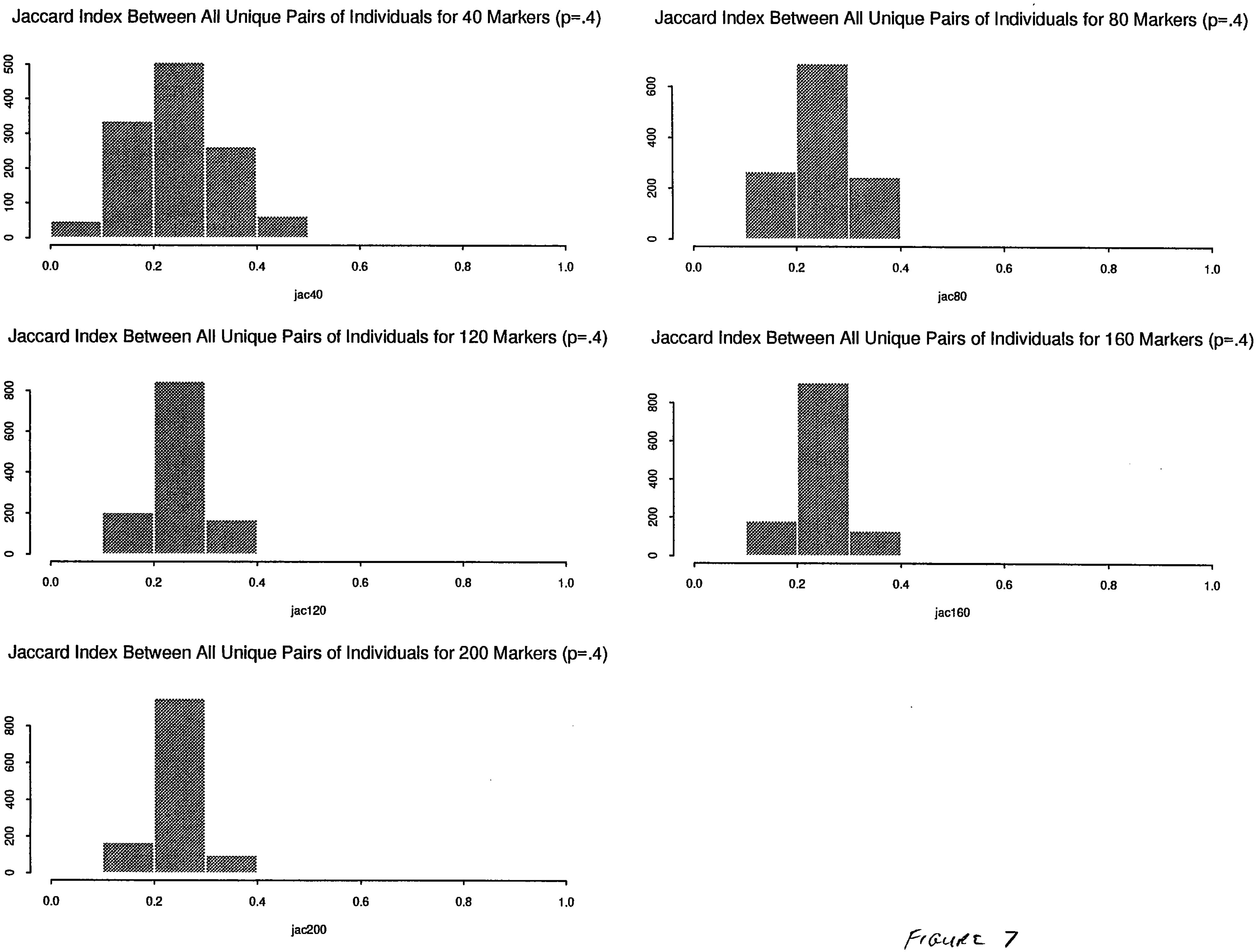
Jaccard Index Between All Unique Pairs of Individuals for 80 Markers (p=.3)

Jaccard Index Between All Unique Pairs of Individuals for 120 Markers (p=.3)

Jaccard Index Between All Unique Pairs of Individuals for 160 Markers (p=.3

Jaccard Index Between All Unique Pairs of Individuals for 200 Markers (p=.3)

FIGURE 6



Jaccard Index Between All Unique Pairs of Individuals for 40 Markers (p=.4)

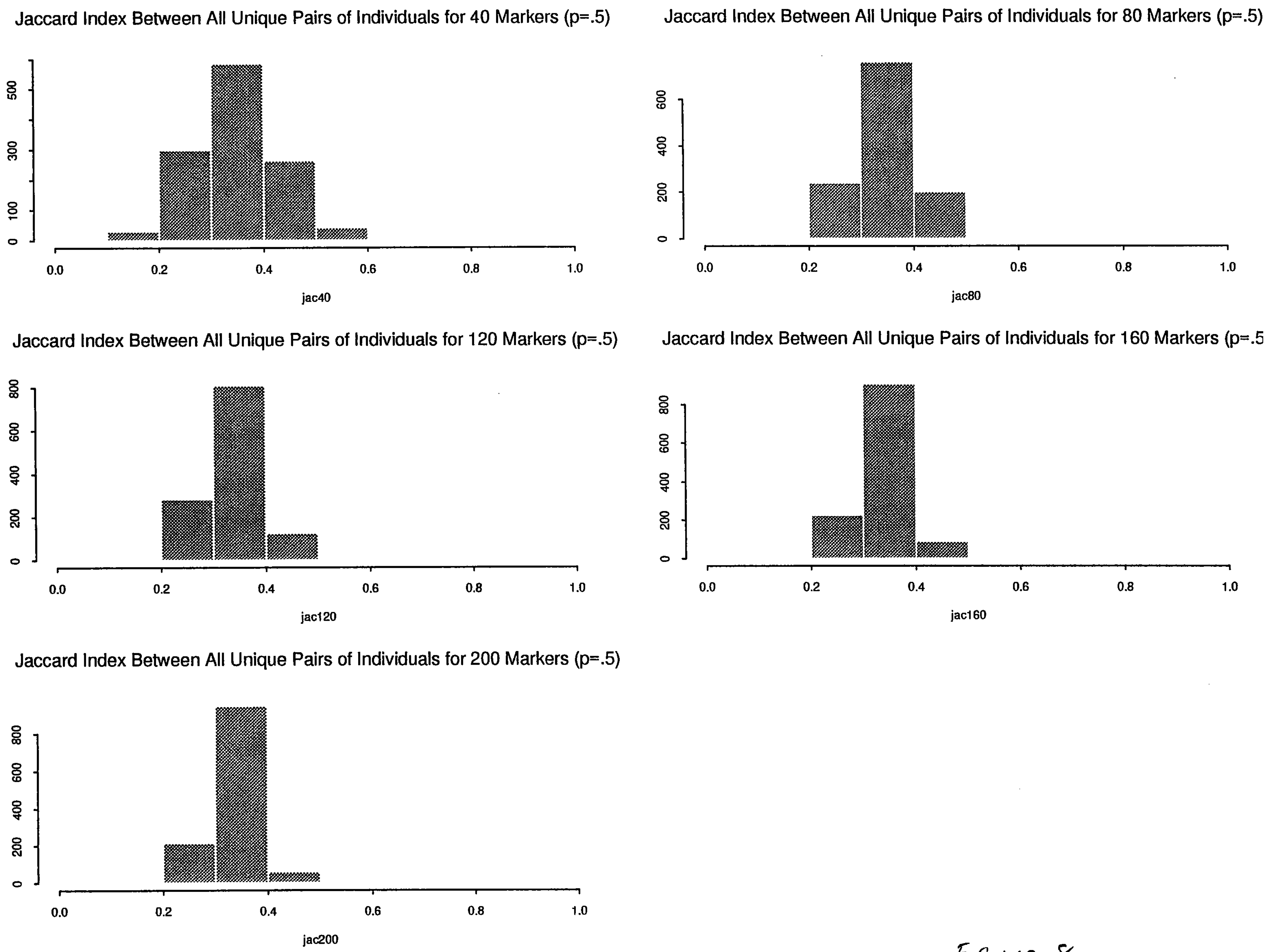
Jaccard Index Between All Unique Pairs of Individuals for 80 Markers (p=.4)

Jaccard Index Between All Unique Pairs of Individuals for 120 Markers (p=.4)

Jaccard Index Between All Unique Pairs of Individuals for 160 Markers (p=.4)

Jaccard Index Between All Unique Pairs of Individuals for 200 Markers (p=.4)

FIGURE 7



Jaccard Index Between All Unique Pairs of Individuals for 40 Markers (p=.5)

Jaccard Index Between All Unique Pairs of Individuals for 80 Markers (p=.5)

Jaccard Index Between All Unique Pairs of Individuals for 120 Markers (p=.5)

Jaccard Index Between All Unique Pairs of Individuals for 160 Markers (p=.5

Jaccard Index Between All Unique Pairs of Individuals for 200 Markers (p=.5)

FIGURE 8

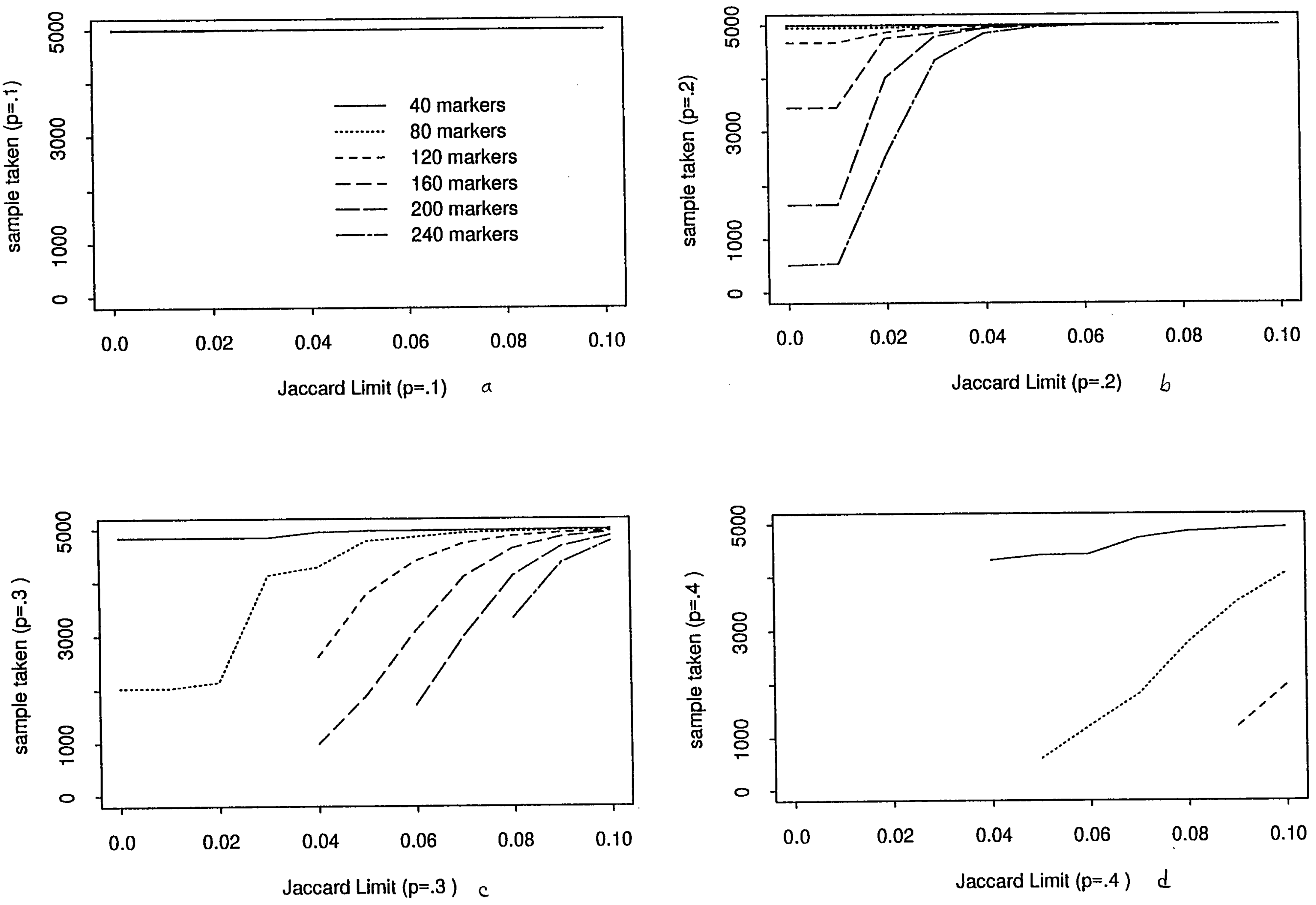

sample taken (p=.1)
5000
3000
1000
0
0.0
0.02
0.04
0.06
0.08
0.10
Jaccard Limit (p=.1)
a
40 markers
80 markers
120 markers
160 markers
200 markers
240 markers
sample taken (p=.2)
5000
3000
1000
0
0.0
0.02
0.04
0.06
0.08
0.10
Jaccard Limit (p=.2)
b
sample taken (p=.3 )
5000
3000
1000
0
0.0
0.02
0.04
0.06
0.08
0.10
Jaccard Limit (p=.3 )
c
sample taken (p=.4 )
5000
3000
1000
0
0.0
0.02
0.04
0.06
0.08
0.10
Jaccard Limit (p=.4 )
d

FIGURE 9

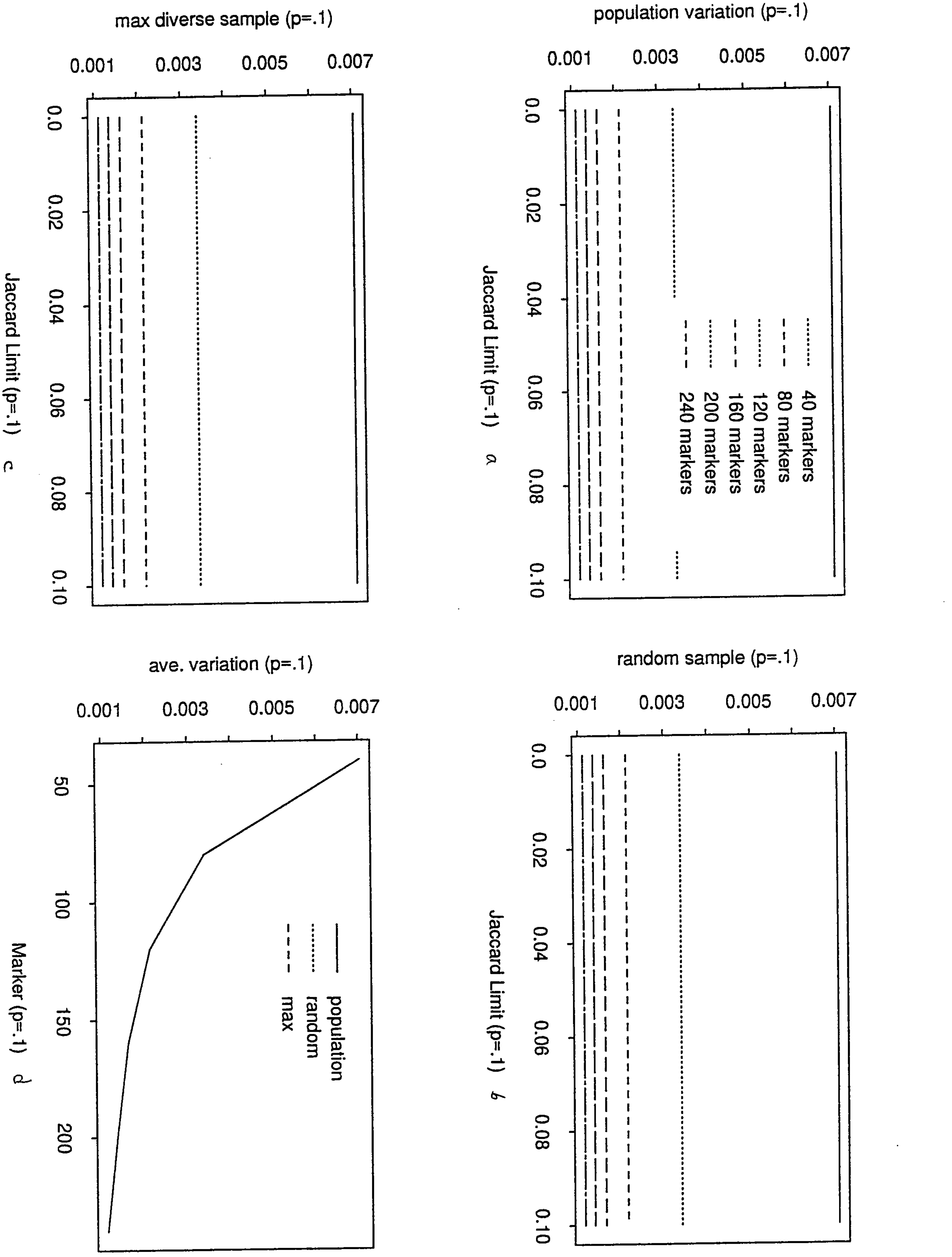

population variation (p=.1)
0.001
0.003
0.005
0.007
0.0
0.02
0.04
0.06
0.08
0.10
40 markers
80 markers
120 markers
160 markers
200 markers
240 markers
Jaccard Limit (p=.1)
a
random sample (p=.1)
0.001
0.003
0.005
0.007
0.0
0.02
0.04
0.06
0.08
0.10
Jaccard Limit (p=.1)
b
max diverse sample (p=.1)
0.001
0.003
0.005
0.007
0.0
0.02
0.04
0.06
0.08
0.10
Jaccard Limit (p=.1)
c
ave. variation (p=.1)
0.001
0.003
0.005
0.007
50
100
150
200
population
random
max
Marker (p=.1)
d

FIGURE 10

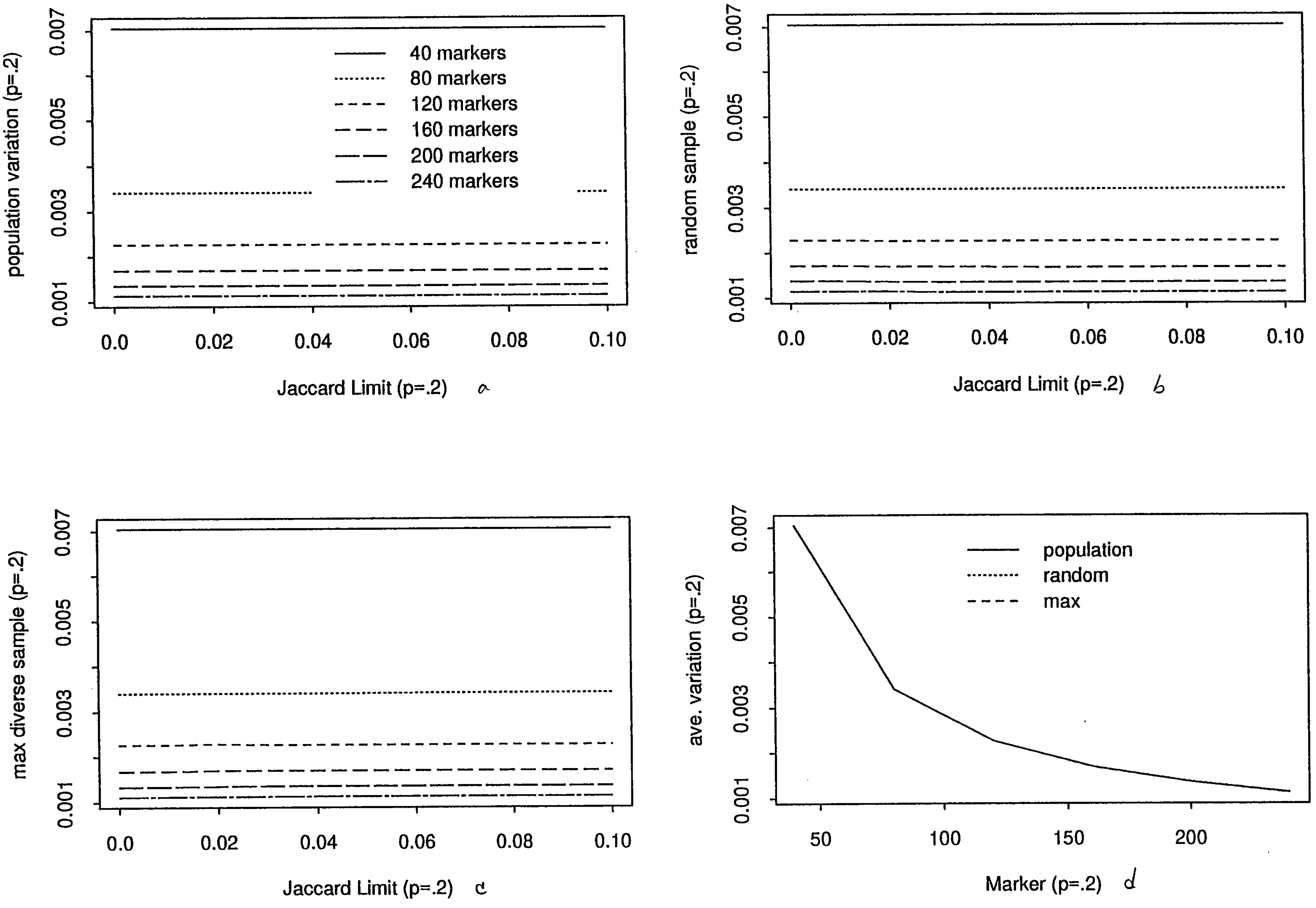
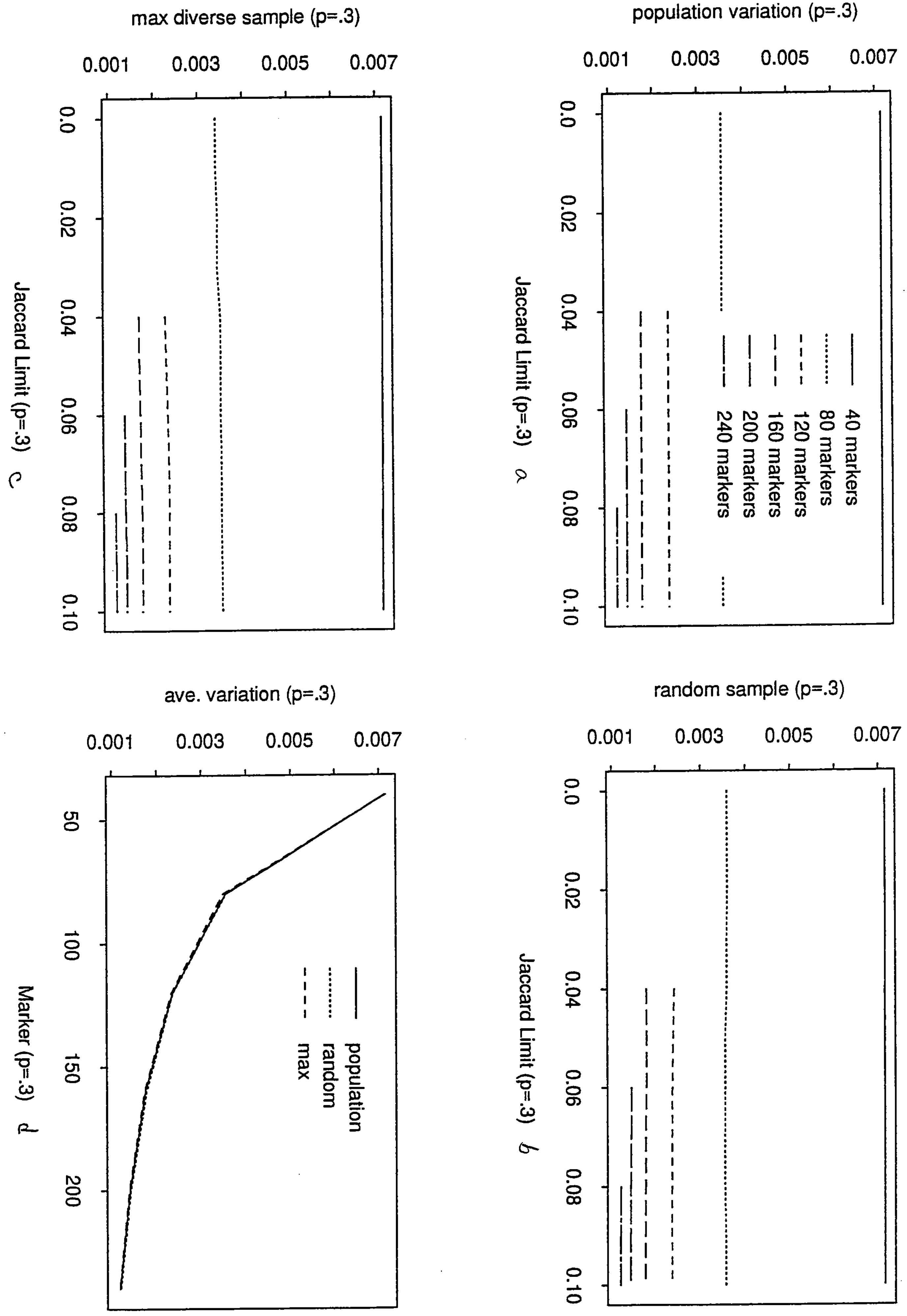

population variation (p=.2)
0.001
0.003
0.005
0.007
0.0
0.02
0.04
0.06
0.08
0.10
Jaccard Limit (p=.2)
a
40 markers
80 markers
120 markers
160 markers
200 markers
240 markers
random sample (p=.2)
0.001
0.003
0.005
0.007
0.0
0.02
0.04
0.06
0.08
0.10
Jaccard Limit (p=.2)
b
max diverse sample (p=.2)
0.001
0.003
0.005
0.007
0.0
0.02
0.04
0.06
0.08
0.10
Jaccard Limit (p=.2)
c
ave. variation (p=.2)
0.001
0.003
0.005
0.007
50
100
150
200
Marker (p=.2)
d
population
random
max

FIGURE 11


population variation (p=.3)
0.001
0.003
0.005
0.007
0.0
0.02
0.04
0.06
0.08
0.10
Jaccard Limit (p=.3)
a
40 markers
80 markers
120 markers
160 markers
200 markers
240 markers
random sample (p=.3)
0.001
0.003
0.005
0.007
0.0
0.02
0.04
0.06
0.08
0.10
Jaccard Limit (p=.3)
b
max diverse sample (p=.3)
0.001
0.003
0.005
0.007
0.0
0.02
0.04
0.06
0.08
0.10
Jaccard Limit (p=.3)
c
ave. variation (p=.3)
0.001
0.003
0.005
0.007
50
100
150
200
Marker (p=.3)
d
population
random
max

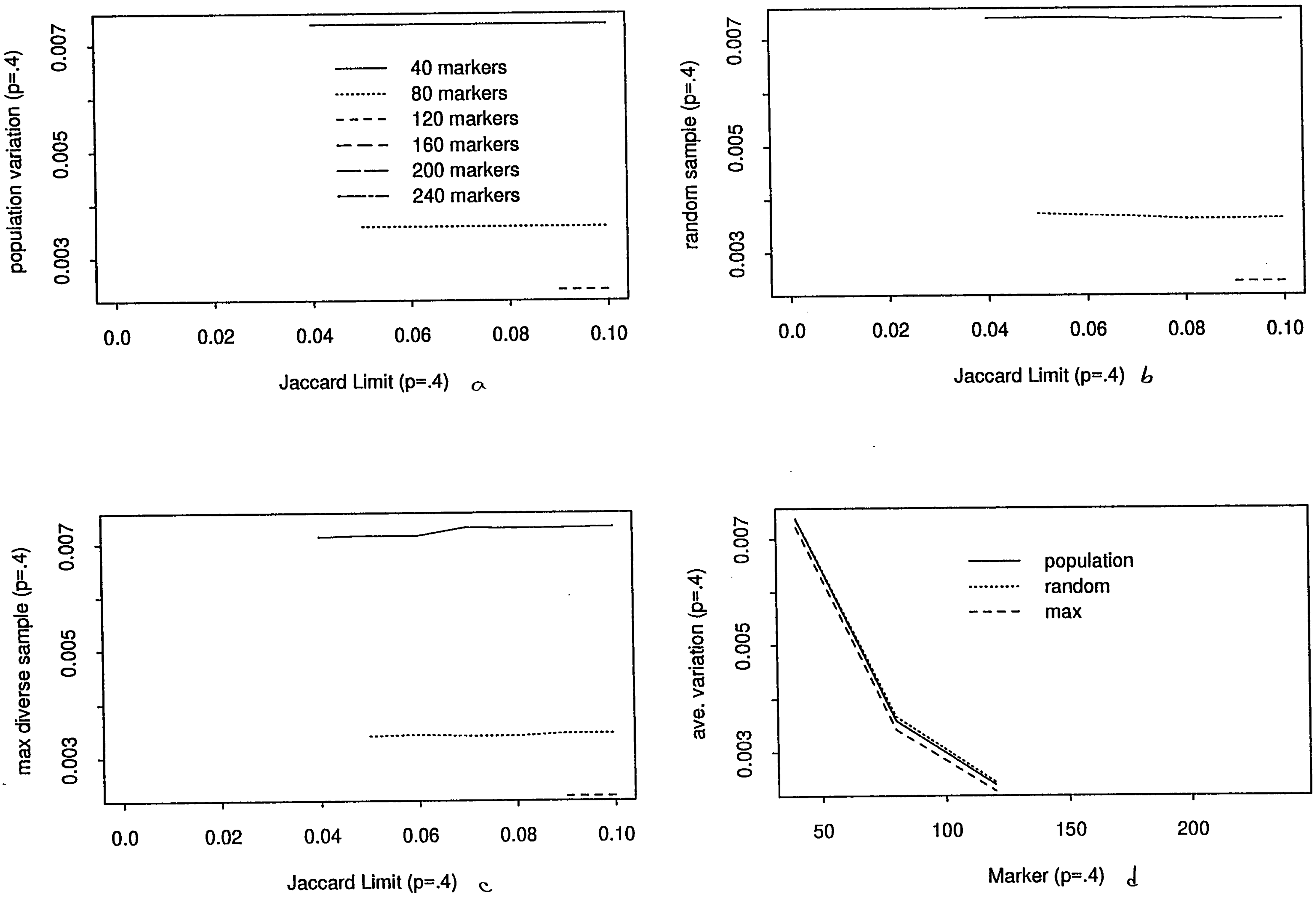

population variation (p=.4)
0.003
0.005
0.007
40 markers
80 markers
120 markers
160 markers
200 markers
240 markers
0.0
0.02
0.04
0.06
0.08
0.10
Jaccard Limit (p=.4)
a
random sample (p=.4)
0.003
0.005
0.007
0.0
0.02
0.04
0.06
0.08
0.10
Jaccard Limit (p=.4)
b
max diverse sample (p=.4)
0.003
0.005
0.007
0.0
0.02
0.04
0.06
0.08
0.10
Jaccard Limit (p=.4)
c
ave. variation (p=.4)
0.003
0.005
0.007
population
random
max
50
100
150
200
Marker (p=.4)
d

FIGURE 13