National Library of Canada
Canadian Theses Division
Ottawa, Canada
K1A 0N4

Bibliothèque nationale du Canada
Division des thèses canadiennes

# PERMISSION TO MICROFILM — AUTORISATION DE MICROFILMER

- Please print or type — Écrire en lettres moulées ou dactylographier

Full Name of Author — Nom complet de l'auteur

Geoffrey Brian Golding

Date of Birth — Date de naissance

July 5 1953

Country of Birth — Lieu de naissance

Canada

Permanent Address — Résidence fixe

National Institute of Environmental Health Sciences
P.O. Box 12233 Research Triangle Park
North Carolina 27709

Title of Thesis — Titre de la thèse

The Application of Identity Coefficients to Problems in Population Genetics

University — Université

University of Alberta

Degree for which thesis was presented — Grade pour lequel cette thèse fut présentée

Ph D.

Year this degree conferred — Année d'obtention de ce grade

1982

Name of Supervisor — Nom du directeur de thèse

C. Strobeck

Permission is hereby granted to the NATIONAL LIBRARY OF CANADA to microfilm this thesis and to lend or sell copies of the film.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

L'autorisation est, par la présente, accordée à la BIBLIOTHÈQUE NATIONALE DU CANADA de microfilmer cette thèse et de prêter ou de vendre des exemplaires du film.

L'auteur se réserve les autres droits de publication; ni la thèse ni de longs extraits de celle-ci ne doivent être imprimés ou autrement reproduits sans l'autorisation écrite de l'auteur.

Date

April 20 1982

Signature

G. Brian Golding

THE UNIVERSITY OF ALBERTA

# THE APPLICATION OF IDENTITY COEFFICIENTS TO PROBLEMS IN POPULATION GENETICS

by

© GEOFFREY BRIAN GOLDING

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE

OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF GENETICS

EDMONTON, ALBERTA

FALL 1982

THE UNIVERSITY OF ALBERTA

RELEASE FORM

NAME OF AUTHOR GEOFFREY BRIAN GOLDING

TITLE OF THESIS THE APPLICATION OF IDENTITY COEFFICIENTS TO PROBLEMS IN POPULATION GENETICS

DEGREE FOR WHICH THESIS WAS PRESENTED DOCTOR OF PHILOSOPHY

YEAR THIS DEGREE GRANTED FALL 1982

Permission is hereby granted to THE UNIVERSITY OF ALBERTA LIBRARY to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

(SIGNED) G. Brian Golding

PERMANENT ADDRESS:

Department of Genetics
University of Alberta
Edmonton, Alberta T6G 2E9

DATED August 20 1982

THE UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research, for acceptance, a thesis entitled THE APPLICATION OF IDENTITY COEFFICIENTS TO PROBLEMS IN POPULATION GENETICS submitted by GEOFFREY BRIAN GOLDING in partial fulfilment of the requirements for the degree of DOCTOR OF PHILOSOPHY.

Curtis Strobeck

Supervisor

K Morgan

C.R. Somerville

External Examiner

Date July 26 1982

To Donna

and to my parents

# ABSTRACT

In this dissertation the method of identity coefficients is used to study several aspects of the genetic structure of populations. This method involves the construction of recursion relationships for the probabilities that a specific sample of gametes are, or are not, identical.

The problems considered are the expected amount of squared linkage disequilibrium between three loci in a random mating population and between two loci in a partially selfing population. The variance of the two-locus, squared linkage disequilibrium in a random mating population is examined. The effects of intragenic recombination among three sites within a gene are determined. The effect of intragenic recombination within a hybrid population and the effect on the variance of homozygosity are examined. The variance of homozygosity within a structured population is derived.

# ACKNOWLEDGEMENTS

I wish to thank C. Strobeck and K. Morgan for their help, guidance and teachings. I wish to express my appreciation to the members of my commitee for their useful comments and for their time. Thanks also go to the staff and students of the Department of Genetics for the environment they have provided. Finally, my thanks go to the Natural Sciences and Engineering Research Council of Canada and to the Alberta Heritage Foundation for Medical Research for their financial support.

# TABLE OF CONTENTS

# LIST OF TABLES

TABLE PAGE

# LIST OF FIGURES

FIGURE PAGE

# Chapter 1

## Introduction

Theoretical population genetics is an attempt to describe mathematically the genetic structures of populations and to determine the consequences of natural processes (such as Mendelian inheritance and selection on such structures). In the following chapters several different problems relevant to theoretical population genetics are examined. Each chapter deals with a different problem.

All of these problems are connected by the method used to solve them. This method has a long history and requires an introduction. In general, the approach is to define a variable, or a set of variables which define the identity relations among a specific set of genes and to determine how these variables change each generation. The resulting system of recursion relationships is at least informative and can often be solved.

While studying the effects of inbreeding on guinea pigs, Wright (1921) found it useful to define a coefficient, f, which describes the degree of inbreeding. This coefficient was defined as the genetic correlation between uniting gametes for a model with two equally frequent alleles. Wright (1922) then extended the method and allowed the correlation f to be defined for all gene frequencies. He called this correlation the inbreeding coefficient. The concept of a variable [illegible] define the effects of inbreeding proved to be very useful. Wright [illegible] was [illegible] Hardy Weinberg equilibrium then the expected frequency of the genotypes AA, Aa and aa would be $p^2 + pqf$, $2pq(1-f)$ and $q^2 + pqf$. Thus, inbreeding increases the frequency of homozygotes and decreases the frequency of heterozygotes. Perhaps based on this simple formula, Wright (1951) changed the definition, notation and name of f. He defined the "inbreeding coefficient (or fixation index) F" as the deviation from Hardy-Weinberg proportions.

Another approach to define the degree of inbreeding was taken by Bernstein (1930), Haldane and Moshinsky (1939), Cotterman (1940) and Malécot (1948). These authors used only probability arguments. Malécot (1948) defined the "coefficient de consanguinité $f_M$" as the probability that the two genes of an individual, M, are identical by descent. This has caused some confusion since f, as a probability, can take values (0, 1) while f, as a correlation, can take values (-1, 1). Furthermore, when the correlation is greater than zero, both indices are equivalent for many models. It is therefore important to realize that these quantities are distinct even though they may have the same values. Unfortunately, the name "inbreeding coefficient" (variously designated f or F) for both indices is now firmly entrenched in the literature (eg: Cavalli-Sforza and Bodmer, 1971; Crow and Kimura, 1970; Jaquard, 1974; Lewontin, 1974; Li, 1978; Malécot, 1969; Roughgarden, 1979; Spiess, 1977; Wright, 1969). In the following, "inbreeding coefficient" will refer to Malécot's definition.

Malécot (1948) extended the method to consider the probability that the genes of two individuals are identical by descent. He defined the "coefficient de parenté $f_{IL}$" as the probability that a gene, chosen at random from individual I, is identical by descent to a gene chosen at random from individual L. This coefficient has also proved to be very useful and is in common use.

A cursory examination of popular texts demonstrates that many different names are used for this coefficient. Malécot (1969) has translated it as the "coefficient of coancestry" and also as the "coefficient of kinship" (from Crow and Kimura, 1970). Crow and Kimura (1970) call it both the "coefficient of kinship" and the "coefficient of consanguinity". Jaquard (1974), Ewens (1979) and Cavalli-Sforza and Bodmer (1971) use the term "coefficient of kinship" while Kempthorne (1957) uses "coefficient of parentage". Falconer (1960) and Roughgarden (1979) prefer "coefficient of coancestry". And finally, Spiess (1977) calls it the "coefficient of transmission". The confusion is added to by Roughgarden's (1979) definition of the "kinship coefficient" as the mean number of alleles identical by descent between two individuals (hence varying from 0 to 2).

These two variables are not, however the final word. The whole concept can be extended quite generally to consider the probabilities of identity among the genes of any particular sample of gametes. Again this has been done independently by several authors.

Harris (1964) described a set of coefficients to study inbreeding and called these "probabilities of alikeness by descent". Cockerham (1971) has also defined a set and simply termed the coefficients as probabilities, or as two, three, etc. "-gene probability functions". Gillois (1964), Jaquard (1974), Chevalet and Gillois (1977, 1978) and Chevalet _et_ _al_. (1977) have defined a set of coefficients to describe the genetic structure of a population. They use the name "coefficients of identity by descent". The inbreeding coefficient and the coefficient of kinship (or whatever is preferred) are relegated to special identity coefficients. Some of these coefficients are identical to Cotterman's "k coefficients". Serant (1974, 1976) defines a set of coefficients to deal with two-locus problems (see also Haldane, 1950 for the effects of inbreeding with linked loci). He calls all of these "inbreeding coefficients" and uses a "kinship process" to derive the recursion relationships. Finally, Cockerham and Weir (1968) define another set of coefficients which they call "descent measures" (Cockerham and Weir, 1973).

Herein, we have followed Jaquard, Gillois and Chevalet and called the variables we are using, "identity coefficients". Again following these authors, we have designated these variables with Greek letters. We do not however, impose the restriction that the relationship among gametes must be "by descent". We consider only their identity by "state", whether or not the gametes carry the same alleles. This is a simpler assumption and allows different (and perhaps easier) formulations to be made for some models.

In the following chapters several different problems are examined to demonstrate the power of this method. Hence several different sets of coefficients are required and are defined in each chapter. The definition of each coefficient is valid only for that chapter. Although a general definition of all coefficients may be possible, this definition would, of necessity, be complicated. As stated above, coefficients can be defined for any desired probability. Therefore, it seemed appropriate to define each coefficient in the context of the problem and to use a notation which imparts as much information as possible in a simple manner. Beyond this, attempts were made to define the coefficients in a consistent way. For other symbols, the notation which is standard in the literature is used.

For the most part, the recursion relationships are placed in appendices. This helps to make the biological relevance of the results clear. Most of these recursion relationships were solved numerically using the IMSL subroutine LEQT1F on an Amdahl 470V/8 in double precision. This subroutine solves the equations using Gaussian elimination (Crout algorithm) with partial pivoting and row scaling. Since the equations solved are approximations, the answers obtained are accurate to the degree that these approximations are correct. To be precise, the relative size of the perturbation in the answers is of the same order of magnitude as the relative size of the perturbation in the system of equations (Noble and Daniel, 1977, Theorem 5.9). The graphs were produced using the University of Alberta's CGPL and CPLT3D subroutines.

All of the problems approached involve a finite population size and fall into two broad categories. The first category deals with linkage disequilibrium (a measure of the correlation between alleles at different loci) in a finite population and in the absence of selection. The second category deals with some effects of intragenic recombination. In chapter 2 it is shown that the effect of recombination within a gene on the expected homozygosity increases when recombination can occur between a larger number of sites. The same model shows that the squared linkage disequilibrium between three loci can be relatively large. In chapter 3 the expected squared linkage disequilibrium, when partial selfing occurs, is determined. It is shown that the linkage disequilibrium can be much larger than that expected for a randomly mating population. The theory developed for randomly mating populations can be applied to a population with partial selfing using a simple transformation which defines "effective" values for the rate of recombination and the population size. Chapter 4 examines the ability of recombination within genes to create unique alleles in hybrids. It is also shown that when such recombination occurs, a randomly mating population may maintain a greater number of alleles than does a structured population. Chapter 5 describes a method to find the variance of homozygosity in a structured population. It is shown that the variance is quite sensitive to the amount of migration. However, the variance can be accurately estimated from the amount of homozygosity. Chapter 6 examines the variance of homozygosity for a gene which consists of two sites, and the variance of squared linkage disequilibrium. It is shown that the coefficient of variation of the linkage disequilibrium is often greater than 100%.

# Chapter 2

# The Expected Variance of Linkage Disequilibrium Between Three Loci in a Finite Population

## Introduction

Since linkage disequilibrium can be generated by epistatic selection between loci, this quantity has been extensively used in the search for the effects of selection in natural populations. Linkage disequilibrium can also be generated by other factors such as recent admixture, assortative mating and random drift in a finite population. Therefore, in order to determine whether the observed values could be caused by random drift, it is necessary to determine the expected value of the variance of linkage disequilibrium in a finite population without selection. This was done by Hill and Robertson (1968) and Ohta and Kimura (1969) for the two-locus model, with two alleles per locus and no mutation, by Weir and Cockerham (1974) for the two-locus model without mutation and by Hill (1975) for the infinite alleles, two-locus model. Experimental studies of linkage disequilibrium, while determining such two-locus linkage disequilibria, often determine the strength of three-locus linkage disequilibria as well (eg: Allard _et al_., 1972; Langley _et al_., 1974; Mukai _et al_., 1974; Brown _et al_., 1977). The study of the expected value of the variance of three-locus linkage disequilibrium in a finite population has been less well developed. Hill (1974a, 1974b) has studied the transient behavior of three-locus linkage disequilibrium with two alleles at each locus and has found that, although three-locus linkage disequilibrium decays faster than two-locus linkage disequilibrium, both can reach appreciable levels before declining to zero.

Here, the expected variance of three-locus linkage disequilibrium is determined in the absence of selection. This is done for the infinite alleles model (Kimura and Crow, 1964), using identity coefficients and then solving the equations numerically on a computer. It is shown that the variance of three-locus linkage disequilibrium is of the same order of magnitude as the variance of two-locus linkage disequilibrium. Hence, even if third order linkage disequilibrium is observed at appreciable values between closely linked loci this is not necessarily an indication of selection. This model can also be interpreted as intragenic recombination between three sites to show that a gene consisting of three sites can have many more alleles present than genes with either one or two sites.

## Theory

The three loci are designated as A, B and C with mutation rates, $\nu_1$ , $\nu_2$ and $\nu_3$, respectively. Let $r_{12}$ be the probability of recombination between A and B, $r_{23}$ between B and C and $r_{13}$ between A and C. It is useful to define $x_1$ as the probability of recombination between B and C but not between A and B, $x_2$ as the probability of recombination between A and B but not between B and C and $x_3$ as the probability of recombination between A and B and between B and C.

Hence

$$x_1 = \tfrac{1}{2}(r_{13}+r_{23}-r_{12})$$
$$x_2 = \tfrac{1}{2}(r_{12}+r_{13}-r_{23})$$
$$x_3 = \tfrac{1}{2}(r_{12}+r_{23}-r_{13})$$

(Strobeck, 1976).

The three-locus model used here assumes a finite population size of 2N gametes and the infinite alleles model of Kimura and Crow (1964). Following the Wright-Fisher model, the chromosomes of the present generation are derived by a random sampling, with replacement, from the chromosomes of the past generation. Each gamete chosen may or may not be a recombinant with the probabilities given above. For example, for any two arbitrary gametes $a_1b_1c_1$ and $a_2b_2c_2$, the meiotic product $a_1b_1c_2$ will be selected with probability $\tfrac{1}{2}x_1$. The sampling process is continued until 2N new gametes have been generated. A similar method was used (Strobeck and Morgan, 1978) to analyze a two site model.

To describe the behavior of the system from one generation to the next requires twenty eight different identity coefficients. Three of these variables define the probability of identity at each of the three loci, A, B and C. As has been shown for the two-locus model, another three variables are required when genes at two loci are considered jointly. Since there are three pairs of loci for the three-locus model (AB, AC and BC), this adds another nine variables. A further sixteen are necessary when the three loci are considered jointly. These coefficients involve choosing two to six distinct gametes at random without replacement. For clarity those which involve two distinct gametes are denoted by $\Phi$, three gametes by $\Gamma$, four gametes by $\Delta$, five gametes by $\Lambda$ and those involving six distinct gametes by $\Psi$. Letters in the subscripts indicate which loci are being considered and a slash is used to seperate those loci which come from different chromosomes. The coefficients are defined in Table 2.1. The sixteen coefficients used by Hill (1974b, Table 3), each a product of gamete frequencies, can be derived from these coefficients by a linear transformation.

If $\Phi_{A/A}$ is the probability that two genes are identical, then $\Phi_{A/A}$ has the recursion relationship

$$\Phi'_{A/A} = (1-\nu_1)^2 \left(\frac{1}{2N} + (1-\frac{1}{2N})\Phi_{A/A}\right)$$

where $\nu_1$ is the mutation rate to neutral, distinct alleles at that locus (Kimura and Crow, 1964). The recursion relationships for two linked loci, ( $\Phi_{AB/AB}$, $\Gamma_{AB/A/B}$ and $\Delta_{A/A/B/B}$ ) are derived by Strobeck and Morgan (1978). Those for $\Phi_{B/B}$, $\Phi_{C/C}$, $\Phi_{BC/BC}$, $\Phi_{AC/AC}$, $\Gamma_{BC/B/C}$, $\Gamma_{AC/A/C}$, $\Delta_{B/B/C/C}$ and $\Delta_{A/A/C/C}$ are the same except that mutation rates and recombination rates have to be changed appropriately. These recursion relationships are included in Appendix 1.

The recursion relationships for three loci are more complicated than those for two loci and it would be very time consuming to write them down. It is therefore useful to make some initial approximations. It is assumed that N >> 1, $\nu_i = O(N^{-1})$ and all of the recombination parameters are of $O(N^{-1})$. Terms with $O(N^{-2})$ are neglected in writing the recursion equations since these terms will affect the answers only

Table 2.1: Definitions of identity coefficients for three linked loci.

$\Phi_{A/A} = \text{Prob}(a_i \equiv a_j)$ $\quad \Phi_{B/B} = \text{Prob}(b_i \equiv b_j)$ $\quad \Phi_{C/C} = \text{Prob}(c_i \equiv c_j)$

$\Phi_{AB/AB} = \text{Prob}(a_i \equiv a_j \text{ and } b_i \equiv b_j)$ $\quad \Phi_{BC/BC} = \text{Prob}(b_i \equiv b_j \text{ and } c_i \equiv c_j)$ $\quad \Phi_{AC/AC} = \text{Prob}(a_i \equiv a_j \text{ and } c_i \equiv c_j)$

$\Gamma_{AB/A/B} = \text{Prob}(a_i \equiv a_j \text{ and } b_i \equiv b_k)$ $\quad \Gamma_{BC/B/C} = \text{Prob}(b_i \equiv b_j \text{ and } c_i \equiv c_k)$ $\quad \Gamma_{AC/A/C} = \text{Prob}(a_i \equiv a_j \text{ and } c_i \equiv c_k)$

$\Delta_{A/A/B/B} = \text{Prob}(a_i \equiv a_j \text{ and } b_k \equiv b_l)$ $\quad \Delta_{B/B/C/C} = \text{Prob}(b_i \equiv b_j \text{ and } c_k \equiv c_l)$ $\quad \Delta_{A/A/C/C} = \text{Prob}(a_i \equiv a_j \text{ and } c_k \equiv c_l)$

$\Phi_{ABC/ABC} = \text{Prob}(a_i \equiv a_j \text{ and } b_i \equiv b_j \text{ and } c_i \equiv c_j)$ $\quad \Gamma_{ABC/AB/C} = \text{Prob}(a_i \equiv a_j \text{ and } b_i \equiv b_j \text{ and } c_i \equiv c_k)$

$\Gamma_{ABC/BC/A} = \text{Prob}(a_i \equiv a_k \text{ and } b_i \equiv b_j \text{ and } c_i \equiv c_j)$ $\quad \Gamma_{ABC/AC/B} = \text{Prob}(a_i \equiv a_j \text{ and } b_i \equiv b_k \text{ and } c_i \equiv c_j)$

$\Gamma_{AB/BC/AC} = \text{Prob}(a_i \equiv a_k \text{ and } b_i \equiv b_j \text{ and } c_j \equiv c_k)$ $\quad \Delta_{ABC/A/B/C} = \text{Prob}(a_i \equiv a_j \text{ and } b_i \equiv b_k \text{ and } c_i \equiv c_l)$

$\Delta_{AB/AB/C/C} = \text{Prob}(a_i \equiv a_j \text{ and } b_i \equiv b_j \text{ and } c_k \equiv c_l)$ $\quad \Delta_{AB/BC/A/C} = \text{Prob}(a_i \equiv a_j \text{ and } b_i \equiv b_k \text{ and } c_k \equiv c_l)$

$\Delta_{AB/AC/B/C} = \text{Prob}(a_i \equiv a_j \text{ and } b_i \equiv b_k \text{ and } c_j \equiv c_l)$ $\quad \Delta_{BC/BC/A/A} = \text{Prob}(a_i \equiv a_j \text{ and } b_k \equiv b_l \text{ and } c_k \equiv c_l)$

$\Delta_{BC/AC/A/B} = \text{Prob}(a_i \equiv a_j \text{ and } b_k \equiv b_l \text{ and } c_k \equiv c_i)$ $\quad \Delta_{AC/AC/B/B} = \text{Prob}(a_i \equiv a_j \text{ and } b_k \equiv b_l \text{ and } c_i \equiv c_j)$

$\Lambda_{AB/A/B/C/C} = \text{Prob}(a_i \equiv a_j \text{ and } b_i \equiv b_k \text{ and } c_l \equiv c_m)$ $\quad \Lambda_{BC/A/A/B/C} = \text{Prob}(a_i \equiv a_j \text{ and } b_k \equiv b_l \text{ and } c_k \equiv c_m)$

$\Lambda_{AC/A/B/B/C} = \text{Prob}(a_i \equiv a_j \text{ and } b_k \equiv b_l \text{ and } c_i \equiv c_m)$ $\quad \Psi_{A/A/B/B/C/C} = \text{Prob}(a_i \equiv a_j \text{ and } b_k \equiv b_l \text{ and } c_m \equiv c_n)$

Where "$\text{Prob}(a_i \equiv a_j)$" is the probability that the allele at locus A from the i-th gamete is identical to the allele at locus A from the j-th gamete.

to $O(N^{-2})$ (Noble and Daniel, 1977, Theorem 5.9). The recursion relationships for the expected values of the 28 coefficients over replicate populations are given in Appendix 1. Additionally let $\nu_1 = \nu_2 = \nu_3 = \nu$, $r_{12} = r_{23} = r$ and $r_{13} = r_{12}+r_{23}$. This is equivalent to a model with complete interference. However, a model without interference implies $r_{13} = r_{12} + r_{23} - 2r_{12}r_{23}$ but by assumption $r_{12}$, $r_{23} << 1$ and thus $r_{13} = r_{12} + r_{23}$. The model therefore holds both with or without interference. Making these substitutions in the equations implies that

$$\Phi_{A/A} = \Phi_{B/B} = \Phi_{C/C}$$

$$\Phi_{AB/AB} = \Phi_{BC/BC}$$

$$\Gamma_{AB/A/B} = \Gamma_{BC/B/C}$$

$$\Delta_{A/A/B/B} = \Delta_{B/B/C/C}$$

$$\Gamma_{ABC/AB/C} = \Gamma_{ABC/BC/A}$$

$$\Delta_{AB/AB/C/C} = \Delta_{BC/BC/A/A}$$

$$\Delta_{AB/AC/B/C} = \Delta_{BC/AC/A/B}$$

$$\Lambda_{AB/A/B/C/C} = \Lambda_{BC/A/A/B/C}$$

which reduces the number of necessary coefficients to nineteen. Even if an explicit equilibrium solution is obtained for each of the coefficients, such solutions would be too complicated to be of value and therefore the equilibrium solutions for various parameter values are obtained numerically on a computer.

Cockerham and Weir (1973), Serant (1974, 1976) have shown that there is a simple relationship between these identity coefficients and gene frequency moments, as is shown in Table 2.2. This table gives the gene frequency moment to which each of the nineteen coefficients

Table 2.2: Relations between the expected values of the identity coefficients over replicate populations and the expected gene frequency moments.

| | | |
|---|---|---|
| $\Phi_{A/A} = E(\sum_{i} p_i^2)$ | $\Phi_{AB/AB} = E(\sum\sum_{ij} f_{ij}^2)$ | $\Phi_{AC/AC} = E(\sum\sum_{ik} f_{i\cdot k}^2)$ |
| $\Gamma_{AB/A/B} = E(\sum\sum_{ij} f_{ij\cdot} p_i q_j)$ | $\Gamma_{AC/A/C} = E(\sum\sum_{ik} f_{i\cdot k} p_i r_k)$ | $\Delta_{A/A/B/B} = E(\sum\sum_{ij} p_i^2 q_j^2)$ |
| $\Delta_{A/A/C/C} = E(\sum\sum_{ik} p_i^2 r_k^2)$ | $\Phi_{ABC/ABC} = E(\sum\sum\sum_{ijk} f_{ijk}^2)$ | $\Gamma_{ABC/AB/C} = E(\sum\sum\sum_{ijk} f_{ijk} f_{ij\cdot} r_k)$ |
| $\Gamma_{ABC/AC/B} = E(\sum\sum\sum_{ijk} f_{ijk} f_{i\cdot k} q_j)$ | $\Gamma_{AB/BC/AC} = E(\sum\sum\sum_{ijk} f_{ij\cdot} f_{\cdot jk} f_{i\cdot k})$ | $\Delta_{ABC/A/B/C} = E(\sum\sum\sum_{ijk} f_{ijk} p_i q_j r_k)$ |
| $\Delta_{AB/AB/C/C} = E(\sum\sum\sum_{ijk} f_{ij\cdot}^2 r_k^2)$ | $\Delta_{AB/BC/A/C} = E(\sum\sum\sum_{ijk} f_{ij\cdot} f_{\cdot jk} p_i r_k)$ | $\Delta_{AB/AC/B/C} = E(\sum\sum\sum_{ijk} f_{ij\cdot} f_{i\cdot k} q_j r_k)$ |
| $\Delta_{AC/AC/B/B} = E(\sum\sum\sum_{ijk} f_{i\cdot k}^2 q_j^2)$ | $\Lambda_{AB/A/B/C/C} = E(\sum\sum\sum_{ijk} f_{ij\cdot} p_i q_j r_k^2)$ | $\Lambda_{AC/A/B/B/C} = E(\sum\sum\sum_{ijk} f_{i\cdot k} p_i q_j^2 r_k)$ |
| $\Psi_{A/A/B/B/C/C} = E(\sum\sum\sum_{ijk} p_i^2 q_j^2 r_k^2)$ | | |

correspond in expectation. Here, $f_{ijk}$ is the frequency of the gamete carrying the i-th allele at locus A, the j-th allele at locus B and the k-th allele at locus C; $f_{ij.}$ is the frequency of the gamete carrying the i-th allele at locus A and the j-th allele at locus B; $p_i$, $q_j$ and $r_k$ are the frequencies of alleles i, j and k at loci A, B and C respectively. The expected linkage disequilibrium, $E(D_{ij})$, between alleles $a_i$ and $b_j$ can be expressed as

$$E(D_{ij}) = E(f_{ij} - p_i q_j)$$

The natural extension of this quantity for three loci is

$$E(D_{ijk}) = E(f_{ijk} - f_{.jk}p_i - f_{i.k}q_j - f_{ij.}r_k + 2p_i q_j r_k)$$

(Hill, 1976). For the model used here both $E(D_{ij})$ and $E(D_{ijk})$ are zero for all i, j and k, however they have non-trivial variances. Hill (1975) determined that at equilibrium

$$E(\sum_{ij}\sum D_{ij}^2) = \frac{16N^2v^2(8Nv+2Nr+5)}{(1+4Nv)(256N^3v^3+192N^3v^2r+32N^3vr^2+320N^2v^2+152N^2vr+8N^2r^2+108Nv+26Nr+9)}$$

With three loci $E(\sum_{ijk}\sum\sum D_{ijk}^2)$ can be approximated as

$$\begin{aligned}E(\sum_{ijk}\sum\sum D_{ijk}^2) = E\Big(\sum_{ijk}\sum\sum\big(f_{ijk}^2 &- 2f_{ijk}f_{ij.}r_k - 2f_{ijk}f_{.jk}p_i - \\ &2f_{ijk}f_{i.k}q_j + 4f_{ijk}p_i q_j r_k + f_{ij.}^2 r_k^2 + \\ &2f_{ij.}f_{.jk}p_i r_k + 2f_{ij.}f_{i.k}q_j r_k - 4f_{ij.}\text{[illegible]} + \\ &f_{.jk}^2 p_i^2 + 2f_{.jk}f_{i.k}p_i q_j - 4f_{.jk}\text{[illegible]} \\ &f_{i.k}^2 q_j^2 - 4f_{i.k}\text{[illegible]}\,r_k + 4\text{[illegible]}\end{aligned}$$

$$= \Phi_{ABC/ABC} - 2\Gamma_{ABC/BC/A} - 2\Gamma_{ABC/AC/B} - 2\Gamma_{ABC/AB/C} +$$
$$4\Delta_{ABC/A/B/C} + \Delta_{BC/BC/A/A} + 2\Delta_{AC/BC/A/B} + 2\Delta_{AB/BC/A/C} -$$
$$4\Lambda_{BC/A/A/B/C} + \Delta_{AC/AC/B/B} + 2\Delta_{AB/AC/B/C} - 4\Lambda_{AC/A/B/B/C} +$$
$$\Delta_{AB/AB/C/C} - 4\Lambda_{AB/A/B/C/C} + 4\Psi_{A/A/B/B/C/C}$$

as long as N is large. Using the values obtained for the identity coefficients, $E(\sum_{ijk}\sum\sum D^2_{ijk})$ can be determined for any values of Nr and Nν.

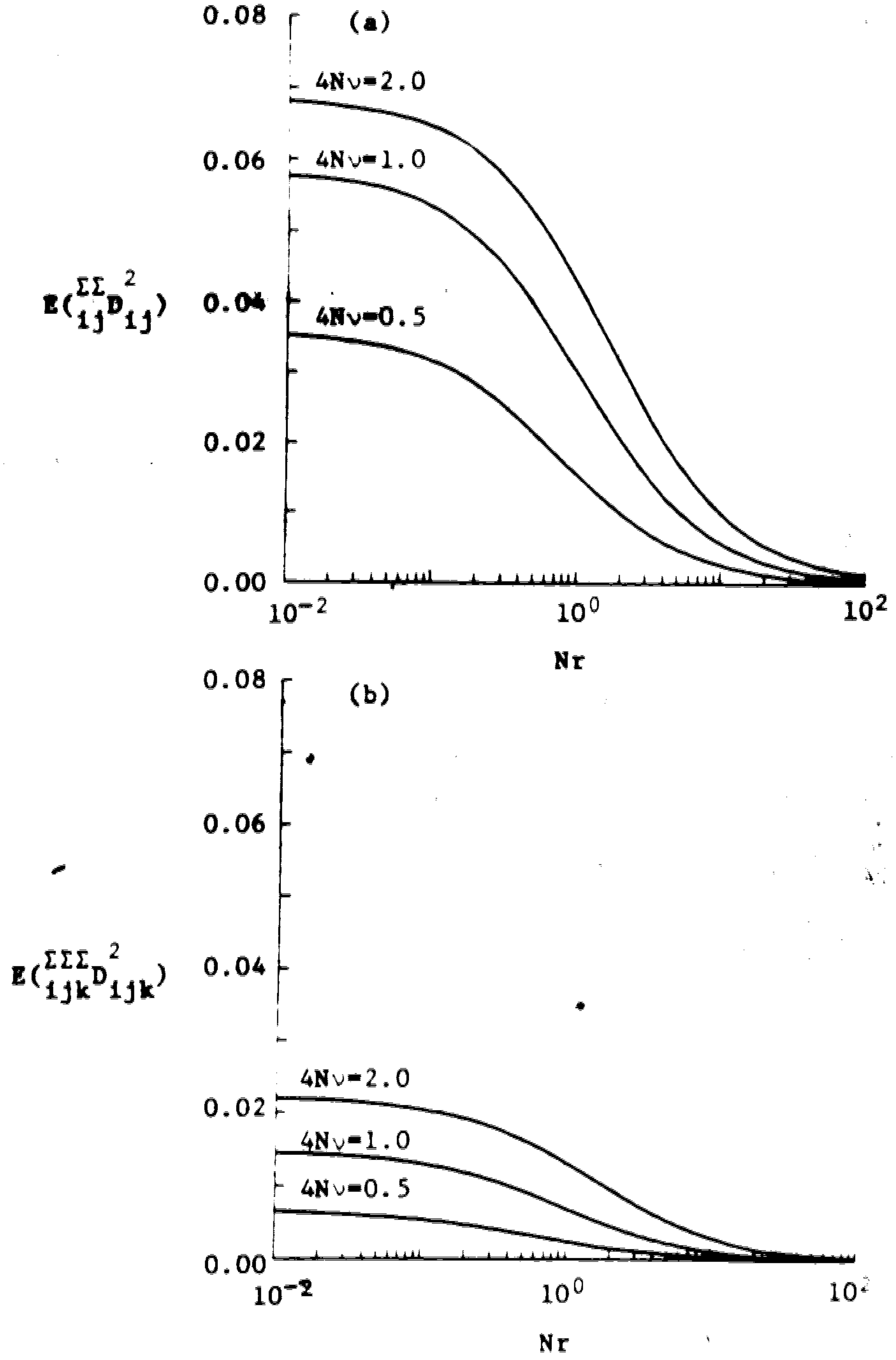
Figure 2.1 gives the value of $E(\sum_{ij}\sum D^2_{ij})$ (2.1a) and $E(\sum_{ijk}\sum\sum D^2_{ijk})$ (2.1b) for 4Nν = 2.0, 1.0 and 0.5 with $10^{-2} \le Nr \le 10^{2}$. It shows that while the variance of three-locus linkage disequilibrium is generally smaller than two-locus they are of the same order of magnitude. Both second and third order linkage disequilibria change values slowly when Nr is less than 0.01 and are negligble when Nr is greater than 10.0. If r = 0 then

$$E(\sum_{ijk}\sum\sum D^2_{ijk}) = \frac{[illegible]\,+130\theta^4+534\theta^3+1074\theta^2+1086\theta+460)}{[illegible]}$$

[illegible] which is thus [illegible] whenever [illegible]

This model can also be interpreted as three sites within a single gene rather than three seperate loci. This interpretation is appropriate for genes which contain introns, a feature of eukaryotic [illegible]

Figure 2.1: The expected squared linkage disequilibrium between two loci, (a) $E(\sum_i\sum_j D_{ij}^2)$, and between three loci, (b) $E(\sum_i\sum_j\sum_k D_{ijk}^2)$.

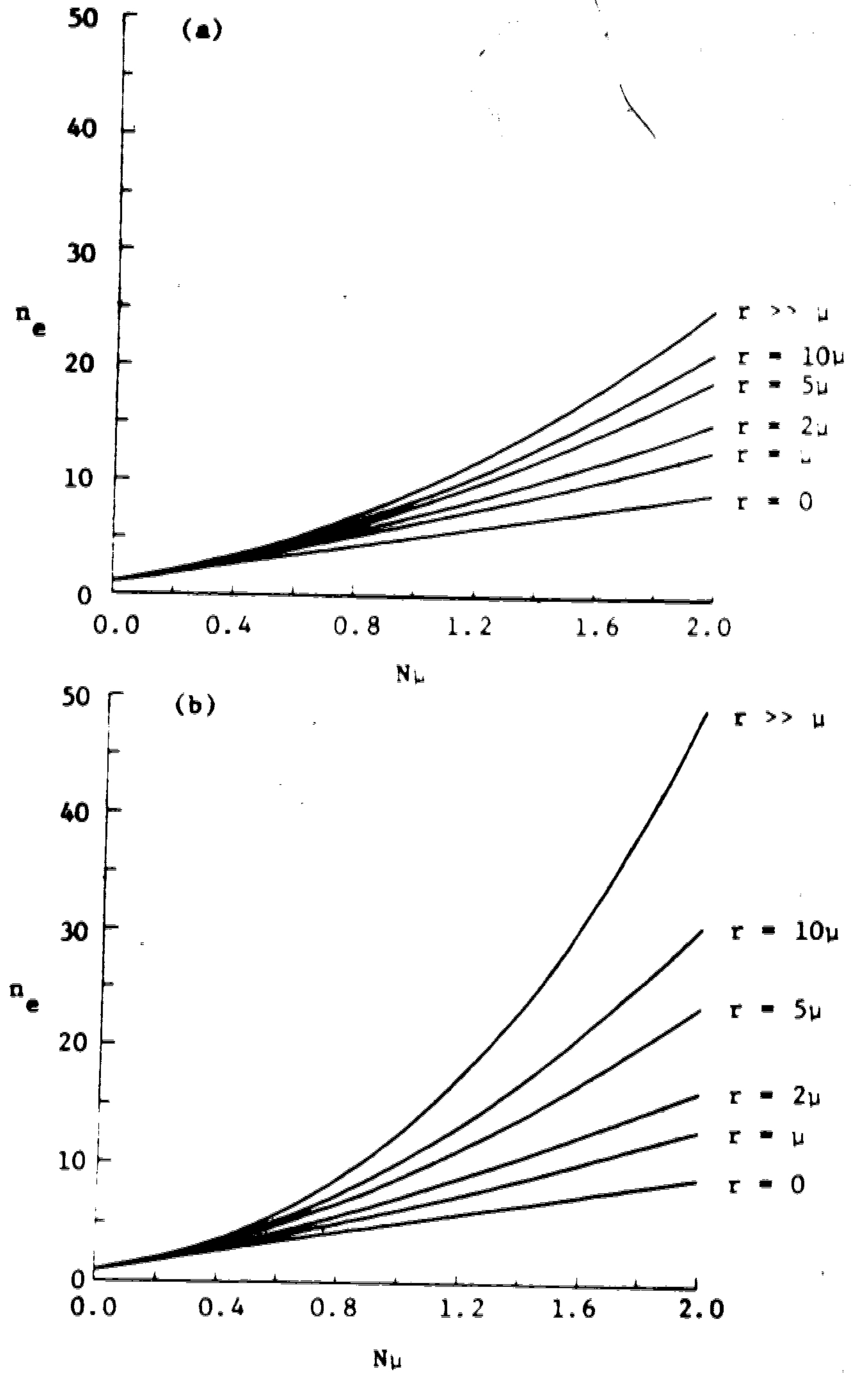

(a)
0.08
0.06
0.04
0.02
0.00
$E(\sum_{ij}\sum D_{ij}^{2})$
4Nν=2.0
4Nν=1.0
4Nν=0.5
$10^{-2}$
$10^{0}$
$10^{2}$
Nr
(b)
0.08
0.06
0.04
0.02
0.00
$E(\sum_{ijk}\sum\sum D_{ijk}^{2})$
4Nν=2.0
4Nν=1.0
4Nν=0.5
$10^{-2}$
$10^{0}$
$10^{2}$
Nr

$\Phi_{ABC/ABC}$ is approximately the expected homozygosity of a single gene with three sites. Figure 2.2 gives the effective number of alleles (one over the homozygosity) for $0 \leq N\mu \leq 2.0$, and $r = 0, \mu, 2\mu, 5\mu, 10\mu$ and $r >> \mu$, where $\mu$ refers to the total mutation rate of the gene ($\mu = 3\nu$ for the three site model) and similarly $r = r_{13} = r_{12}+r_{23}$ is the recombination rate within the whole gene. The effective number of alleles is larger when the gene consists of three sites than when the gene consists of two sites. However, a comparison of Figure 2.2a taken from Strobeck and Morgan (1978, Figure 1) with Figure 2.2b shows that the two site model is a good approximation of the three site model if $r < 2\mu$.

## Summary

The variance of three-locus linkage disequilibria for an equilibrium infinite alleles model is solved numerically on a computer, using identity coefficients. It is shown that the variance of three-locus linkage disequilibrium created by random drift, although smaller than the variance of two-locus linkage disequilibrium, is of the same order of magnitude. Hence third order disequilibria are not necessarily good indications of selection. The formula for the variance of linkage disequilibrium is given when there is no recombination between the genes. This model can also be interpreted as intragenic recombination between three sites within a gene.

Figure 2.2: The effective number of alleles for a gene consisting of two sites (a), and of three sites (b).


(a)
50
40
30
20
10
0
$n_e$
0.0
0.4
0.8
1.2
1.6
2.0
Nμ
r >> μ
r = 10μ
r = 5μ
r = 2μ
r = μ
r = 0
(b)
50
40
30
20
10
0
$n_e$
0.0
0.4
0.8
1.2
1.6
2.0
Nμ
r >> μ
r = 10μ
r = 5μ
r = 2μ
r = μ
r = 0

# Chapter 3

# Linkage Disequilibrium in a Finite Population that is Partially Selfing

## Introduction

There have been several studies on the amount of linkage disequilibrium found in natural populations. Most of these studies found no significant linkage disequilibrium between loci that are not associated with an inversion (Lewontin, 1974; Langley, Ito and Voelker, 1977). However, in plant populations that are partially selfing, a significant amount of linkage disequilibrium is consistently present (Brown, 1979). This observed linkage disequilibrium could be generated either by selection with epistatic interactions between the loci or by random drift. In order to determine whether or not this observed disequilibrium could be a result of random drift, it is necessary to know the amount of linkage disequilibrium expected in a partially selfing finite population without selection.

The expected amount of linkage disequilibrium in a finite population with random mating has been studied extensively. These studies have assumed two alleles at each locus with no mutation (Hill and Robertson, 1968; Ohta and Kimura, 1969), a two-locus model with no mutation (Weir and Cockerham, 1974) or an infinite number of alleles at each locus with mutant alleles differing from all pre-existing ones (Hill, 1975) i.e., the infinite-allele model of Kimura and Crow (1964). In this chapter, the amount of linkage disequilibrium expected in a finite population assuming the infinite allele model and partial selfing is derived using identity coefficients. It is shown that the formulas for the expected sum of squares of the linkage disequilibria and the squared standard linkage disequilibrium are equivalent to those from random mating with a reduced recombination value and a reduced population size.

## Theory

Before considering random drift of two loci in a finite population that is primarily selfing, the one-locus model is developed.

Let the population consist of N diploid individuals that produce offspring by both selfing and outcrossing. Let S be the proportion of the offspring of an individual that are produced by selfing and 1-S the proportion of offspring produced by outcrossing. Each of the N individuals in the next generation is the offspring of either one individual selected at random (if it is produced by selfing) or two individuals selected at random without replacement (if it is produced by outcrossing) from the present generation. If S = 1/N, then there is random union of gametes.

Two identity coefficients are needed to describe the behavior of the system from one generation to the next. One coefficient, $\Psi_{(A/A)}$, is the probability that the two genes of an individual are identical (Malécot's inbreeding coefficient). The other coefficient, $\Phi_{(A)(A)}$, is the probability that two genes selected from two different individuals are identical (Malécot's kinship coefficient). (The notation used for the subscripts is explained when considering the two-locus model.) Since the probability of an offspring having its two genes identical is $1/2 + 1/2\Psi_{(A/A)}$ if it is produced by selfing and $\Phi_{(A)(A)}$ if it is produced by outcrossing,

$$\Psi_{(A/A)}' = (1-\mu)^2\{S(\tfrac{1}{2}+\tfrac{1}{2}\Psi_{(A/A)}) + (1-S)\Phi_{(A)(A)}\} \qquad (1a)$$

$$\Phi_{(A)(A)}' = (1-\mu)^2\{\tfrac{1}{N}(\tfrac{1}{2}+\tfrac{1}{2}\Psi_{(A/A)}) + (1-\tfrac{1}{N})\Phi_{(A)(A)}\} \qquad (1b)$$

where $\mu$ is the mutation rate to unique alleles. These relations define the expected value of the coefficients in the next generation in terms of their previous values.

If $N >> 1$ and $\mu = O(\frac{1}{N})$, then these equations can be approximated by

$$\Psi_{(A/A)}' = S(\tfrac{1}{2}+\tfrac{1}{2}\Psi_{(A/A)}) + (1-S)\Phi_{(A)(A)} \qquad (2a)$$

if terms of $O(\frac{1}{N})$ or less are neglected, and

$$\Phi_{(A)(A)}' = \tfrac{1}{N}(\tfrac{1}{2}+\tfrac{1}{2}\Psi_{(A/A)}) + (1-\tfrac{1}{N} - 2\mu)\Phi_{(A)(A)} \qquad (2b)$$

if terms of $O(\frac{1}{N^2})$ or less are neglected. At equilibrium

$$\Psi_{(A/A)} = \frac{\tfrac{1}{2}S + (1-S)\Phi_{(A)(A)}}{1 - \tfrac{1}{2}S}$$

from (2a) and substituting this value into (2b)

$$\hat{\Phi}_{(A)(A)} = \frac{1}{1 + 4N\mu - 2N\mu S} = \frac{1}{1 + 4N_e\mu} \qquad (3a)$$

and therefore

$$\hat{\Psi}_{(A/A)} = \frac{1 + 2N\mu S}{1 + 4N\mu - 2N\mu S} = \frac{1 + 2N\mu S}{1 + 4N_e\mu} \qquad (3b)$$

where $N_e = (1-\frac{1}{2}S)N$. It can be verified that these are the approximate equilibrium values of equations (1a) and (1b) by substitution or from the theory of perturbed matrices (section 5.5, Noble and Daniel, 1977). If $\Phi_A$ is the probability that two genes chosen randomly from the population without replacement (not necessarily from two different individuals) are identical, then

$$\Phi_A = \frac{1}{2N-1}\Psi_{(A/A)} + (1-\frac{1}{2N-1})\Phi_{(A)(A)} \simeq \Phi_{(A)(A)}$$

since $N >> 1$ (Cockerham, 1967).

We now turn our attention to the two-locus model. Denote the two loci by A and B, and let r be the recombination value between them. Let N be the number of diploid individals, S be the proportion of selfing and $\mu$ and $\nu$ be the mutation rates to unique alleles at the A and B loci, respectively.

Sixteen identity coefficients are required to describe random drift of two loci in a finite populatin that is partially selfing. These identity coefficients involve randomly choosing chromosomes without replacement from one, two, three or four different individuals and are denoted by $\Psi$, $\Phi$, $\Gamma$ and $\Delta$ respectively. The following notation is used in the subscripts: parentheses are used to separate the genes contributed by different individuals, and slashes are used to separate the genes contributed by different chromosomes of an individual. For example, $\Phi_{(AB)(A/B)}$ is the probability of identity at both loci if the genes at the A and B loci are chosen from one chromosome of one inidivual and from different chromosomes of another individual. If the genes on the two chromosomes of an arbitrary individual are denoted by $a_{i1}b_{i1}$ and $a_{i2}b_{i2}$, respectively, then the sixteen identity coefficients are given in Table 3.1.

The recursion relationships for the expected values of the sixteen identity coefficients over replicate populations are given in Appendix 2. At equilibrium,

$$
\begin{aligned}
\bar{\Psi}_{(A/A)} &= \frac{\tfrac{1}{2}S + (1-S)\bar{\Phi}_{(A)(A)}}{1 - \tfrac{1}{2}S} \\
\bar{\Psi}_{(B/B)} &= \frac{\tfrac{1}{2}S + (1-S)\bar{\Phi}_{(B)(B)}}{1 - \tfrac{1}{2}S} \\
\bar{\Psi}_{(AB/AB)} &= \frac{\tfrac{1}{2}S + (1-S)\bar{\Phi}_{(AB)(AB)}}{1 - \tfrac{1}{2}S} \\
\bar{\Phi}_{(AB)(A/B)} &= \frac{\tfrac{1}{2}S\bar{\Phi}_{(AB)(AB)} + (1-S)\bar{\Gamma}_{(AB)(A)(B)}}{1 - \tfrac{1}{2}S} \\
\bar{\Phi}_{(AB/B)(A)} &= \frac{\tfrac{1}{2}\bar{\Phi}_{(A)(A)} + (1-S)\bar{\Gamma}_{(AB)(A)(B)}}{1 - \tfrac{1}{2}S} \\
\bar{\Phi}_{(AB/A)(B)} &= \frac{\tfrac{1}{2}\bar{\Phi}_{(B)(B)} + (1-S)\bar{\Gamma}_{(AB)(A)(B)}}{1 - \tfrac{1}{2}S} \\
\bar{\Gamma}_{(B/B)(A)(A)} &= \frac{\tfrac{1}{2}S\bar{\Phi}_{(A)(A} + (1-S)\bar{\Delta}_{(A)(B)(A)(B)}}{- \tfrac{1}{2}S} \\
\bar{\Gamma}_{(A/A)(B)(B)} &= \frac{\tfrac{1}{2}S\bar{\Phi}_{(B)(B)} + (1-S)\bar{\Delta}_{(A)(B)(A)(B)}}{1 - \tfrac{1}{2}S}
\end{aligned}
\qquad (4)
$$

Table 3.1: Definitions of identity coefficients for a partial selfing population.

$$\Psi_{(A/A)} = P(a_{i1} \equiv a_{i2}) \qquad \Psi_{(B/B)} = P(b_{i1} \equiv b_{i2})$$

$$\Phi_{(A)(A)} = P(a_{i1} \equiv a_{j1}) \qquad \Phi_{(B)(B)} = P(b_{i1} \equiv b_{j1})$$

$$\Psi_{(AB/AB)} = P(a_{i1} \equiv a_{i2} \text{ and } b_{i1} \equiv b_{i2})$$

$$\Phi_{(AB)(AB)} = P(a_{i1} \equiv a_{j1} \text{ and } b_{i1} \equiv b_{j1})$$

$$\Phi_{(AB)(A/B)} = P(a_{i1} \equiv a_{j1} \text{ and } b_{i1} \equiv b_{j2})$$

$$\Phi_{(AB/B)(A)} = P(a_{i1} \equiv a_{j1} \text{ and } b_{i1} \equiv b_{i2})$$

$$\Phi_{(AB/A)(B)} = P(a_{i1} \equiv a_{i2} \text{ and } b_{i1} \equiv b_{j1})$$

$$\Phi_{(A/A)(B/B)} = P(a_{i1} \equiv a_{i2} \text{ and } b_{j1} \equiv b_{j2})$$

$$\Phi_{(A/B)(A/B)} = P(a_{i1} \equiv a_{j2} \text{ and } b_{i1} \equiv b_{j2})$$

$$\Gamma_{(AB)(A)(B)} = P(a_{i1} \equiv a_{j1} \text{ and } b_{i1} \equiv b_{k1})$$

$$\Gamma_{(B/B)(A)(A)} = P(a_{j1} \equiv a_{k1} \text{ and } b_{i1} \equiv b_{i2})$$

$$\Gamma_{(A/A)(B)(B)} = P(a_{i1} \equiv a_{i2} \text{ and } b_{j1} \equiv b_{k1})$$

$$\Gamma_{(A/B)(A)(B)} = P(a_{i1} \equiv a_{j1} \text{ and } b_{i2} \equiv b_{k1})$$

$$\Delta_{(A)(B)(A)(B)} = P(a_{i1} \equiv a_{k1} \text{ and } b_{j1} \equiv b_{l1})$$

The genes of the two chromosomes of an individual are denoted by $a_{i1}b_{i1}$ and $a_{i2}b_{i2}$, respectively. ("$\equiv$" is read "is identical to").

$$\hat{\Gamma}_{(A/B)(A)(B)} = \frac{\frac{1}{2}S\hat{\Gamma}_{(AB)(A)(B)} + (1-S)\hat{\Delta}_{(A)(B)(A)(B)}}{1 - \frac{1}{2}S}$$

$$\hat{\Phi}_{(A/A)(B/B)} = \frac{\frac{1}{2}S^2(1+\frac{1}{2}S) + \frac{3}{4}S^2(1-S)(\hat{\Phi}_{(A)(A)}+\hat{\Phi}_{(B)(B)}) + (1-S)^2(1+\frac{3}{2}S)\hat{\Delta}_{(A)(B)(A)(B)}}{(1+\frac{1}{2}S)(1-\frac{1}{2}S)^2}$$

$$\hat{\Phi}_{(A/B)(A/B)} = \frac{\frac{1}{2}S^2(1+\frac{1}{2}S)\hat{\Phi}_{(AB)(AB)} + \frac{3}{2}S^2(1-S)\hat{\Gamma}_{(AB)(A)(B)} + (1-S)^2(1+\frac{3}{2}S)\hat{\Delta}_{(A)(B)(A)(B)}}{(1+\frac{1}{2}S)(1-\frac{1}{2}S)^2}$$

from (A1). Substituting these values into (A2) gives the equilibrium values of the identity coefficients $\Phi_{(A)(A)}$, $\Phi_{(B)(B)}$, $\Phi_{(AB)(AB)}$, $\Gamma_{(AB)(A)(B)}$ and $\Delta_{(A)(B)(A)(B)}$ as shown in Table 3.2, where

$$U = N(1-\tfrac{1}{2}S)\mu = N_e\mu$$

$$V = N(1-\tfrac{1}{2}S)\nu = N_e\nu$$

$$R = N(1-S)r = N(1-\tfrac{1}{2}S)(1-S)r/(1-\tfrac{1}{2}S) = N_e r_e$$

The equilibrium values of the other identity coefficients are obtained by substituting the equilibrium values from Table 3.2 into (4).

In order to compare these results for a partially selfing population to the equivalent results for a random mating population, it is necessary to define five further identity coefficients. Three of these identity coefficients involve choosing two chromosomes at random without replacement from the population. One coefficient involves choosing three chromosomes and one coefficient involves choosing four chromosomes. (The chromosomes are not necessarily from different individuals). If an arbitrary chromosome is denoted by $a_i b_i$ then the five identity coefficients are

$$\Phi_A = P(a_1 \equiv a_2)$$

3.2: Expected equilibrium values of the identity [illegible]icients for a finite, partially selfing population.

$$[illegible]_{(A)(A)} = \frac{1}{1+4U}$$

$$[illegible]_{(B)(B)} = \frac{1}{1+4V}$$

$$[illegible]_{(AB)(AB)} =$$

$$\frac{[illegible]\,18(U+V)+2R+9]+32(U+V)^3+48(U+V)^2R+16(U+V)R^2+80(U+V)^2+76(U+V)R+8R^2+54(U+V)+26R+9}{[illegible]\,1+4[illegible])[32(U+V)^3+48(U+V)^2R+16(U+V)R^2+80(U+V)^2+76(U+V)R+8R^2+54(U+V)+26R+9]}$$

$$\frac{[illegible]UV[2(U+V)+3]+32(U+V)^3+48(U+V)^2R+16(U+V)R^2+80(U+V)^2+76(U+V)R+8R^2+54(U+V)+26R+9}{[illegible]1+4U)(1+4V)[32(U+V)^3+48(U+V)^2R+16(U+V)R^2+80(U+V)^2+76(U+V)R+8R^2+54(U+V)+26R+9]}$$

$$\frac{32UV+32(U+V)^3+48(U+V)^2R+16(U+V)R^2+80(U+V)^2+76(U+V)R+8R^2+54(U+V)+26R+9}{[illegible]V)[32(U+V)^3+48(U+V)^2R+16(U+V)R^2+80(U+V)^2+76(U+V)R+8R^2+54(U+V)+26R+9]}$$

$$\phi_B = P(b_1 \equiv b_2)$$

$$\phi_{AB} = P(a_1 \equiv a_2 \text{ and } b_1 \equiv b_2)$$

$$\Gamma_{AB} = P(a_1 \equiv a_2 \text{ and } b_1 \equiv b_3)$$

$$\Delta_{AB} = P(a_1 \equiv a_3 \text{ and } b_2 \equiv b_4)$$

(Strobeck and Morgan, 1978). In terms of the previous sixteen identity coefficients,

$$\phi_A = \frac{1}{2N-1}\gamma_{(A/A)} + (1-\frac{1}{2N-1})\phi_{(A)(A)} = \phi_{(A)(A)}$$

$$\phi_B = \frac{1}{2N-1}\gamma_{(B/B)} + (1-\frac{1}{2N-1})\phi_{(B)(B)} = \phi_{(B)(B)}$$

$$\phi_{AB} = \frac{1}{2N-1}\gamma_{(AB/AB)} + (1-\frac{1}{2N-1})\phi_{(AB)(AB)} = \phi_{(AB)(AB)}$$

$$\Gamma_{AB} = \frac{1}{2N}(\phi_{AB\ A\ B} + \phi_{AB\ A\ B} + \phi_{AB\ B\ A}) + \frac{2N-4}{2N}\ [illegible]_{AB\ A\ B} = [illegible]_{AB\ A\ (B)}$$

$$\Delta_{AB} = [illegible]\ \phi_{A\ A\ B\ B} + 2\phi_{A\ B\ A\ B} + \frac{2N-4}{[illegible]}\ [illegible] = [illegible]_{B\ A\ A}$$

[illegible]

[illegible] coefficients are as given in Table 3.2 and are identical to t[illegible] obtained assuming random mating with a population size N = [illegible] and a recombination value r = [illegible] (Strobeck and Mor[illegible] [illegible]). Therefore, the effect of partial selfing at equilibrium is [illegible] population size [illegible]

[illegible]

variation of linkage disequilibrium expected in a finite population (Serant, 1976; Strobeck and Morgan, 1978). If $p_i$ is the frequency of the i-th allele $a_i$ at the A locus, $q_j$ the frequency of the j-th allele $b_j$ at the B locus, and $f_{ij} = p_iq_j + D_{ij}$ the frequency of the chromosome $a_ib_j$, where $D_{ij}$ is the linkage disequilibrium between $a_i$ and $b_j$, then the expected sum of squares of the linkage disequilibria

$$E(\sum_{ij}\sum D_{ij}^2) = \frac{16UV[2(U+V)+1][4(U+V)+2R+5]}{(1+4U)(1+4V)[32(U+V)^3+48(U+V)^2R+16(U+V)R^2+80(U+V)^2+76(U+V)R+8R^2+54(U+V)+26R+9]}$$

and the squared standard linkage disequilibrium

$$\sigma_d^2 = \frac{E(\sum_{ij}\sum D_{ij}^2)}{E(\sum\limits_{\substack{i,k\\ i\neq k}} \sum\limits_{\substack{j,l\\ j\neq l}} p_ip_kq_jq_l)} = \frac{4(U+V)+2R+5}{16(U+V)^2+24(U+V)R+8R^2+32(U+V)+26R+11}$$

(Hill, 1975). In Figures 3.1 and 3.2, the equilibrium values of $E(\sum_{ij}\sum D_{ij}^2)$ and $\sigma_d^2$ are plotted for $10^{-3} \leq Nr \leq 10^3$ and with Nu = Nv = 0.25 and 1.0 and S = 0.0, 0.5, 0.9, 0.99 and 1.0. It is seen that $E(\sum_{ij}\sum D_{ij}^2)$ and $\sigma_d^2$ remain significantly greater than zero for increasingly larger values of Nr as S approaches one and are not functions of the recombination value if S = 1. If r = 0, $E(\sum_{ij}\sum D_{ij}^2)$ has a maximum value when U = V [illegible] 0.505, whereas $\sigma_d^2$ is a decreasing function of U + V. Therefore, increasing the proportion of selfing decreases the value of $\sigma_d^2$, but may increase or decrease $E(\sum_{ij}\sum D_{ij}^2)$ when [illegible]. Thus, the squared standard linkage disequilibrium [illegible] better measure of the variation of linkage [illegible]

Figure 3.1: The expected value of the squared linkage disequilibrium for $N\mu = N\nu = 0.25$ and 1.0 and with partial selfing at a rate S = 0.0, 0.5, 0.9, 0.99, and 1.0 (– – S = 0.0, - - S = 0.5, ‾ ‾ S = 0.9, – – – S = 0.99, —— S = 1.0).

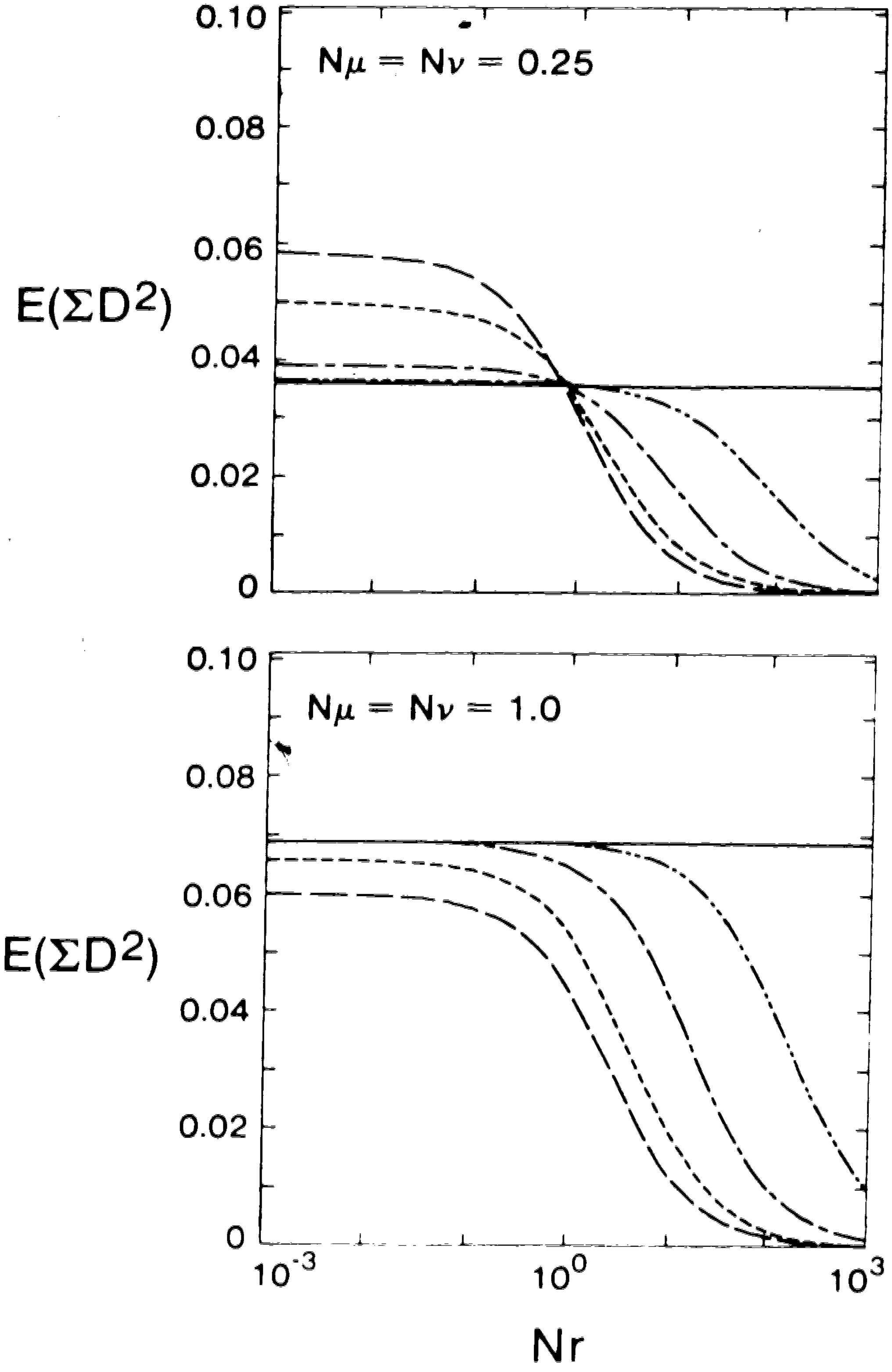
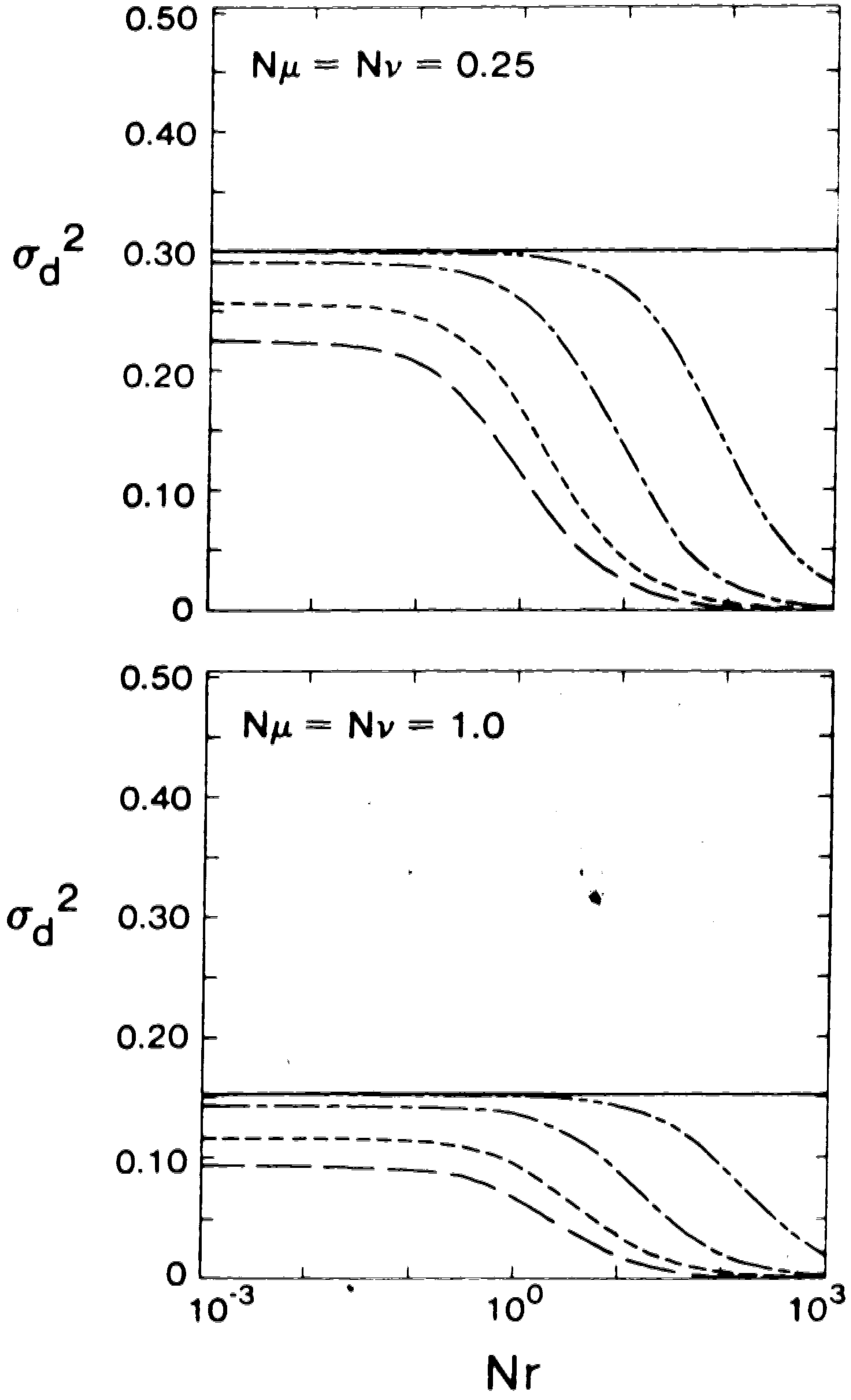

Nμ = Nν = 0.25
Nμ = Nν = 1.0
E(ΣD2)
E(ΣD2)
0.10
0.08
0.06
0.04
0.02
0
0.10
0.08
0.06
0.04
0.02
0
10-3
100
103
Nr

Figure 3.2: The expected value of the squared standard linkage disequilibrium for $N\mu = N\nu = 0.25$ and 1.0 and with partial selfing at a rate $S = 0.0$, 0.5, 0.9, 0.99, and 1.0 (– – $S = 0.0$, - - $S = 0.5$, – - $S = 0.9$, – - - $S = 0.99$, —— $S = 1.0$).


0.50
0.40
0.30
0.20
0.10
0
$\sigma_d^2$
$N\mu = N\nu = 0.25$
0.50
0.40
0.30
0.20
0.10
0
$\sigma_d^2$
$N\mu = N\nu = 1.0$
$10^{-3}$
$10^0$
$10^3$
Nr

## Discussion

The results in the previous section show that there is significant variance in the expected linkage disequilibrium due to random drift in a partially selfing population if

$$N_e r_e = N(1-S)r \leq 1$$

and the mutation rates $\mu$ and $\nu$ are of the order 1/N. It is, therefore, appropriate to examine the experimental data collected from populations of partially selfing plants to see if the observed linkage disequilibrium can be explained by mutation and random drift. The magnitude of (1-S)r will be used as an indicator of whether the observed linkage disequilibrium could be due to random drift. Since the mutation rate is generally assumed to be between $10^{-4}$ and $10^{-8}$, the population size must be larger than approximately $10^{4}$ if the variation is to be maintained in the population. Therefore, (1-S)r must be less than $10^{-4}$ before the observed linkage disequilibrium is likely to be the result of random drift.

In barley, Horedeum vulgare, Allard and his co-workers (Allard, Kahler and Weir, 1972; Weir, Allard and Kahler, 1972, 1974) found significant linkage disequilibrium between four esterase loci in Composite Cross V. Three loci, A, B and C, are closely linked and the fourth locus is unlinked to the other three. The recombination value between the three linked loci are estimated to be $r_{AB}$ = 0.0023, $r_{AC}$ = 0.0048 and $r_{BC}$ = 0.0059 Kahler and [illegible] [illegible]. The estimate of the proportion of selfing is S = 0.9943 (Allard, Kahler and Weir, 1972). Therefore, the value of (1-S)r between AB, AC and BC are 0.000013, 0.000027 and 0.000034, respectively. These values are in the range such that linkage disequilibria could be generated by random drift. However, since Composite Cross V was initiated in 1941, a transient analysis is more appropriate than the comparison of the observed sum of squares of the linkage disequilibria or the squared standard linkage disequilibrium to that expected at equilibrium.

Also, the linkage disequilibrium between six loci, four esterase loci $E_1$, $E_4$, $E_9$ and $E_{10}$, a phosphatase $P_5$, and an anodal peroxidase $APX_5$, has been analyzed in <u>Avena</u> <u>barbata</u>, the slender wild oat, by Allard <u>et</u> <u>al</u>. (1972). Three loci, $P_5$,$APX_5$ and $E_{10}$ are linked, and the recombination values are $r_{P_5-APX_5}$ = 0.04, $r_{APX_5-E_{10}}$ = 0.23 and $r_{P_5-E_{10}}$ = 0.25 (Marshall and Allard, 1969). The proportion of selfing has been estimated to be approximately S = 0.98 (Marshall and Allard, 1970; Hamrick and Allard, 1972). Therefore, the smallest value of (1-S)r, which is between $P_5$ and $APX_5$, is 0.0008. This value is small enough that random drift might have a significant effect if the size of the effective population is relatively small, but the actual population size was estimated to be approximately 50,000.

These two examples show that random drift might explain some of the linkage disequilibrium observed in natural populations. However, random drift is unlikely to be the cause of the observed linkage disequilibrium between [illegible] linked

## Summary

The variation of linkage disequilibrium expected in a finite, partially selfing population is analyzed, assuming the infinite allele model. Formulas for the expected sum of squares of the linkage disequilibria and the squared standard linkage disequilibrium are derived from the equilibrium values of sixteen identity coefficients required to describe the behavior of the system. These formulas are identical to those obtained with random mating if the effective population size

$$N_e = (1- \tfrac{1}{2}S)N$$

and the effective recombination value

$$r_e = (1-S)r/(1- \tfrac{1}{2}S)$$

where S is the proportion of selfing, are substituted for the population size and the recombination value. Therefore, the effect of partial selfing at equilibrium is to reduce the population size by a factor 1- ½S and the recombination value by a factor (1-S)/(1- ½S).

# Chapter 4

## Increased Effective Number of Alleles Found in Hybrid Populations due to Intragenic Recombination

### Introduction

Hybridization is recognized as a common feature of many natural populations and has numerous implications in the study of speciation. One unusual feature of hybrid populations is the presence of alleles which do not exist in either of the parental populations. These unique alleles have been observed by Hunt and Selander (1973) in *Mus musculus musculus* and *M. m. domesticus* hybrids and by Sage and Selander (1979) in *Rana berlandieri* and *R. utriculata* hybrids. Two explanations have been proposed to explain their presence: These unique alleles may be due to increased mutation rates in hybrids (Thompson and Woodruff, 1978) or they may be the result of intragenic recombination between different alleles of the parental populations (Watt, 1972). Ohno *et al.* (1969), McCarron *et al.* (1974), Freeling (1976), Koehn and Eanes (1976), Morgan and Strobeck (1979), and Tsuno (1981) have observed patterns of variability which they attribute to intragenic recombination.

In order to determine if intragenic recombination can explain the presence of these rare alleles, we have constructed a model to determine the amount of variability expected in a finite hybrid [illegible] is assumed to consist of individuals from each parental population which can mate either with individuals from the same or from the opposite parental population. Therefore, the model used is one with two semi-isolated populations which exchange a proportion of their genes each generation. To allow for the possibility of intragenic recombination, the genes are assumed to consist of two sites or parts. It has been shown that both intragenic recombination and population subdivision can increase the variability maintained in a finite population. Intragenic recombination significantly increases the effective number of alleles whenever $Nr > 1$ and $r > \mu$ (where N is the population size, $\mu$ is the mutation rate to neutral alleles and r is the recombination rate between two sites within the gene) (Strobeck and Morgan, 1978). Subdivision of a population can also increase the effective number of alleles in the total population (of size 2N) because a different group of alleles is maintained in each of the subpopulations, although each subpopulation (of size N) has reduced variability (Malécot, 1948). Nei and Feldman (1972) and Chakraborty and Nei (1974) have also studied gene differentiation and rates of change of homozygosity in a subdivided population (for a review see Felsenstein, 1976; Maruyama, 1977).

It is shown here that, at equilibrium, the combination of intragenic recombination and population subdivision increases the effective number of alleles maintained in a population beyond the sum of the effects of each process alone. This effect is greatest when the recombination value is large and hybridization (migration) occurs at an intermediate rate. The transient behavior of the system shows that sympatry of two previously isolated populations can increase the effective number of alleles maintained in each population and in the hybrid population above their equilibrium values for long periods of time after the beginning of hybridization. These results imply that intragenic recombination may be the cause of the observed unique alleles in hybrid populations.

## Theory

Let the population consist of two semi-isolated subpopulations with $N_1$ and $N_2$ diploid individuals. Each generation, the i-th subpopulation receives a proportion $1-m_i$ of its genes from itself and a proportion $m_i$ from the other subpopulation. The gametic migration considered here is equivalent to individual migration for the parameter values of interest to this study. Each gene is assumed to consist of two sites or parts, denoted a and b, which recombine with a probability of r. Both site a and site b can mutate to unique, selectively neutral forms or "alleles" (as in the infinite alleles model of Kimura and Crow, 1964). Let this mutation rate per gamete per generation be $\nu_1$ for site a and $\nu_2$ for site b. Therefore, the mutation rate of the gene is $\mu = \nu_1 + \nu_2$ per gamete per generation.

Malecot (1948), Maruyama (1970) and Nei and Feldman (1972) showed that the behavior of a single locus in such a subdivided population can be described by three identity coefficients (denoted by $\Psi_{(a)_1(a)_1}$, $\Psi_{(a)_1(a)_2}$ and $\Psi_{(a)_2(a)_2}$). $\Psi_{(a)_i(a)_i}$ is defined as the probability that two genes, chosen randomly without replacement from subpopulation i, are identical (that is, both genes carry the same allele) and $\Psi_{(a)_1(a)_2}$ is defined as the probability that a gene chosen from subpopulation 1 is identical to a gene chosen from subpopulation 2. The recursion relationships for the expected values of $\Psi_{(a)_1(a)_1}$, $\Psi_{(a)_1(a)_2}$ and $\Psi_{(a)_2(a)_2}$ over replicate populations are

$$\Psi'_{(a)_1(a)_1} = (1-\nu_1)^2\Big((1-m_1)^2[\frac{1}{2N_1} + (1-\frac{1}{2N_1})\Psi_{(a)_1(a)_1}] + 2m_1(1-m_1)\Psi_{(a)_1(a)_2} + m_1^2[\frac{1}{2N_2} + (1-\frac{1}{2N_2})\Psi_{(a)_2(a)_2}]\Big)$$

$$\Psi'_{(a)_1(a)_2} = (1-\nu_1)^2\Big((1-m_1)(1-m_2)\Psi_{(a)_1(a)_2} + m_1(1-m_2)[\frac{1}{2N_2} + (1-\frac{1}{2N_2})\Psi_{(a)_2(a)_2}]$$
$$+ m_2(1-m_1)[\frac{1}{2N_1} + (1-\frac{1}{2N_1})\Psi_{(a)_1(a)_1}] + m_1m_2\Psi_{(a)_1(a)_2}\Big)$$

$$\Psi'_{(a)_2(a)_2} = (1-\nu_1)^2\Big((1-m_2)^2[\frac{1}{2N_2} + (1-\frac{1}{2N_2})\Psi_{(a)_2(a)_2}] + 2m_2(1-m_2)\Psi_{(a)_1(a)_2} + m_2^2[\frac{1}{2N_1} + (1-\frac{1}{2N_1})\Psi_{(a)_1(a)_1}]\Big)$$

(where the ' indicates the value of the coefficient in the next generation). These same relationships hold true for a single site within a gene and those for a second site, ( $\Psi_{(b)_1(b)_1}$, $\Psi_{(b)_1(b)_2}$ and $\Psi_{(b)_2(b)_2}$), are also the same but with $\nu_1$ replaced by $\nu_2$.

This approach, using identity coefficients, can be extended to consider two linked sites (within a gene) in two subpopulations. Although genes actually consist of many sites the consideration of two sites is an appropriate model to indicate if intragenic recombination has any significant effect. The widespread occurence of introns in eukaryotic genes (Gilbert, 1978; Crick, 1979) facilitates recombination between exons and this model is a fairly accurate representation of a gene with a single intron (the sites being identified with the exons).

To describe the behavior of this system from one generation to the next requires 26 identity coefficients (each the probability that a particular sample of genes, picked at random without replacement, are identical at the a and/or b sites). The symbol $\Psi$ is used to designate the probability that two gametes have identical a (or b) sites. The symbols $\Phi$, $\Gamma$ and $\Delta$ are used to designate probabilities of identity at both the a and b sites chosen from two, three and four gametes respectively. Each symbol is subscripted to indicate how the gametes are chosen. The coefficient $\Phi_{(ab)_1(ab)_2}$, for example, is defined as

$$\Phi_{(ab)_1(ab)_2} = \text{Prob}(\ a_{i1} \equiv a_{j2} \text{ and } b_{i1} \equiv b_{j2}\ )$$

where an arbitrary gamete chosen from the first subpopulation is denoted by $a_{i1}b_{i1}$, an arbitrary gamete chosen from the second subpopulation is denoted by $a_{j2}b_{j2}$ and where "$\equiv$" should be read "is identical to". The definitions of the 26 coefficients are given in Table 4.1.

Complete recursion relationships for these 26 coefficients have been derived and if $N_1$ [illegible] $_i$, $r, m_i = O(1/N_i)$ and terms of [illegible] are neglected the recursion relationships for the expected values of the coefficients over replicate populations simplify to the form given in Appendix 4. Equations similar to those given here were developed by [illegible] (1974) with $N_1 = N_2$ although no analysis of the equations was presented.

[illegible] identity coefficients for two loci

| | | |
|---|---|---|
| [illegible] | $\psi_{(a)_1(a)_2}$ = Prob($a_{i1}{:}a_{j2}$) | $\psi_{(a)_2(a)_2}$ = Prob($a_{i2}{:}a_{j2}$) |
| [illegible] | $\psi_{(b)_1(b)_2}$ = Prob($b_{i1}$ $b_{j2}$) | $\psi_{(b)_2(b)_2}$ = Prob($b_{i2}{:}b_{j2}$) |
| Prob(a [illegible] a [illegible] and $b_{i1}$ $b_{j1}$) | $\phi_{(ab)_1(ab)_2}$ = Prob($a_{i1}{:}a_{j2}$ and $b_{i1}{:}b_{j2}$) | $\phi_{(ab)_2(ab)_2}$ = Prob($a_{i2}{:}a_{j2}$ and $b_{i2}{:}b_{j2}$) |
| Prob(a [illegible] $a_{j1}$ and [illegible] $b_{k1}$) | $\Gamma_{(ab)_1(a)_1(b)_2}$ = Prob($a_{i1}{:}a_{j1}$ and $b_{i1}{:}b_{k2}$) | $\Gamma_{(ab)_1(a)_2(b)_1}$ = Prob($a_{i1}{:}a_{j2}$ and $b_{i1}{:}b_{k1}$) |
| Prob(a [illegible] $a_{j1}$ and $b_{i2}$ $b_{k1}$) | $\Gamma_{(ab)_1(a)_2(b)_2}$ = Prob($a_{i1}{:}a_{j2}$ and $b_{i1}{:}b_{k2}$) | $\Gamma_{(ab)_2(a)_1(b)_2}$ = Prob($a_{i2}$ $a_{j1}$ and $b_{i2}{:}b_{k2}$) |
| Prob($a_{i2}$ $a_{j2}$ and $b_{i2}$ $b_{k1}$) | $\Gamma_{(ab)_2(a)_2(b)_2}$ = Prob($a_{i2}{:}a_{j2}$ and $b_{i2}{:}b_{k2}$) | |
| Prob($a_{i1}$ $a_{j1}$ and $b_{k1}$ $b_{l1}$) | $\Delta_{(a)_1(b)_1(a)_1(b)_2}$ = Prob($a_{i1}{:}a_{j1}$ and $b_{k1}{:}b_{l2}$) | $\Delta_{(a)_1(b)_1(a)_2(b)_1}$ = Prob($a_{i1}{:}a_{j2}$ and $b_{k1}{:}b_{l1}$) |
| Prob(a [illegible] $a_{j2}$ and $b_{k1}$ $b_{l2}$) | $\Delta_{(a)_1(b)_2(a)_1(b)_2}$ = Prob($a_{i1}{:}a_{j1}$ and $b_{k2}{:}b_{l2}$) | $\Delta_{(a)_2(b)_1(a)_2(b)_1}$ = Prob($a_{i2}{:}a_{j2}$ and $b_{k1}{:}b_{l1}$) |
| [illegible] $a_{j2}$ and $b_{k2}$ $b_{l2}$) | $\Delta_{(a)_2(b)_1(a)_2(b)_2}$ = Prot($a_{i2}{:}a_{j2}$ and $b_{k1}{:}b_{l2}$) | $\Delta_{(a)_2(b)_2(a)_2(b)_2}$ = Prob($a_{i2}{:}a_{j2}$ and $b_{k2}{:}b_{l2}$) |

"$a_{in}$ $a_{jm}$" signifies that the a site from the i-th [illegible] in the n-th population is identical to the a site [illegible] in the m-th population.

transient values of homozygosity for a single locus with two populations of size $N_1$ and $N_2$ differ little from those with two populations each of size $N = (N_1+N_2)/2$. We also find that solutions to the equations are not changed qualitatively when it is assumed that $N_1 = N_2 = N$, $\nu_1 = \nu_2 = \nu$ and $m_1 = m_2 = m$. These assumptions greatly simplify the equations at equilibrium since not all 26 coefficients are then required. The number of necessary coefficients is reduced to 11 since

$$\begin{gathered}
\Psi_{(a)_1(a)_1} = \Psi_{(a)_2(a)_2} = \Psi_{(b)_1(b)_1} = \Psi_{(b)_2(b)_2} \\
\Psi_{(a)_1(a)_2} = \Psi_{(b)_1(b)_2} \\
\Phi_{(ab)_1(ab)_1} = \Phi_{(ab)_2(ab)_2} \\
\Gamma_{(ab)_1(a)_1(b)_1} = \Gamma_{(ab)_2(a)_2(b)_2} \\
\Gamma_{(ab)_1(a)_1(b)_2} = \Gamma_{(ab)_1(a)_2(b)_1} = \Gamma_{(ab)_2(a)_2(b)_1} = \Gamma_{(ab)_2(a)_1(b)_2} \\
\Gamma_{(ab)_1(a)_2(b)_2} = \Gamma_{(ab)_2(a)_1(b)_1} \\
\Delta_{(a)_1(b)_1(a)_1(b)_1} = \Delta_{(a)_2(b)_2(a)_2(b)_2} \\
\Delta_{(a)_1(b)_1(a)_1(b)_2} = \Delta_{(a)_1(b)_1(a)_2(b)_1} = \Delta_{(a)_1(b)_2(a)_2(b)_2} = \Delta_{(a)_2(b)_1(a)_2(b)_2} \\
\Delta_{(a)_1(b)_2(a)_1(b)_2} = \Delta_{(a)_2(b)_1(a)_2(b)_1}
\end{gathered}$$

Eleven coefficients were used to determine the equilibrium values of the identity coefficients. If the coefficients initially satisfy the above equalities and the above assumptions are met then the coefficients will satisfy these equalities for all time. Therefore, only these 11 coefficients were also used to study the transient behavior. The equilibrium identity coefficients were found by solving the system of 11 linear equations with particular values of the mutation, migration and recombination rates. The values of these [illegible]

[illegible]

## Results and Discussion

The coefficient $\Phi_{(ab)_1(ab)_1}$ is the expected homozygosity in subpopulation 1 of a gene consisting of two sites (each with a mutation rate $\nu$). The expected homozygosity in subpopulation 2, ($\Phi_{(ab)_2(ab)_2}$) is the same as in subpopulation 1 because of the assumptions of equal mutation rates, migration rates and population sizes. The effective number of alleles, a measure of variability, within each subpopulation is therefore $n_e = 1/\Phi_{(ab)_1(ab)_1}$. If the subdivision is known to an observer, $n_e$ is the appropriate measure of variability in a subdivided population. However, in many cases the population may be spatially or ethologically subdivided and not recognized as such by an observer. In this case genes would be sampled randomly from each group and the expected homozygosity would be

$$\frac{N_1(N_1-1)}{(N_1+N_2)(N_1+N_2-1)}\Phi_{(ab)_1(ab)_1} + \frac{2N_1N_2}{(N_1+N_2)(N_1+N_2-1)}\Phi_{(ab)_1(ab)_2} + \frac{N_2(N_2-1)}{(N_1+N_2)(N_1+N_2-1)}\Phi_{(ab)_2(ab)_2}$$

and the appropriate effective number of alleles would be

$$\frac{N_1 N_1 \text{[illegible]}}{(N_1+N_2)(N_1+N_2-1)}\Phi_{(ab)_1(ab)_1} + \frac{2N_1N_2}{(N_1+N_2)(N_1+N_2-1)}\Phi_{(ab)_1(ab)_2} + \frac{N_2(N_2 \text{[illegible]}}{(N_1+N_2)(N_1+N_2 \text{[illegible]}}$$

since $N_1, N_2 \gg 1$. With the above assumptions

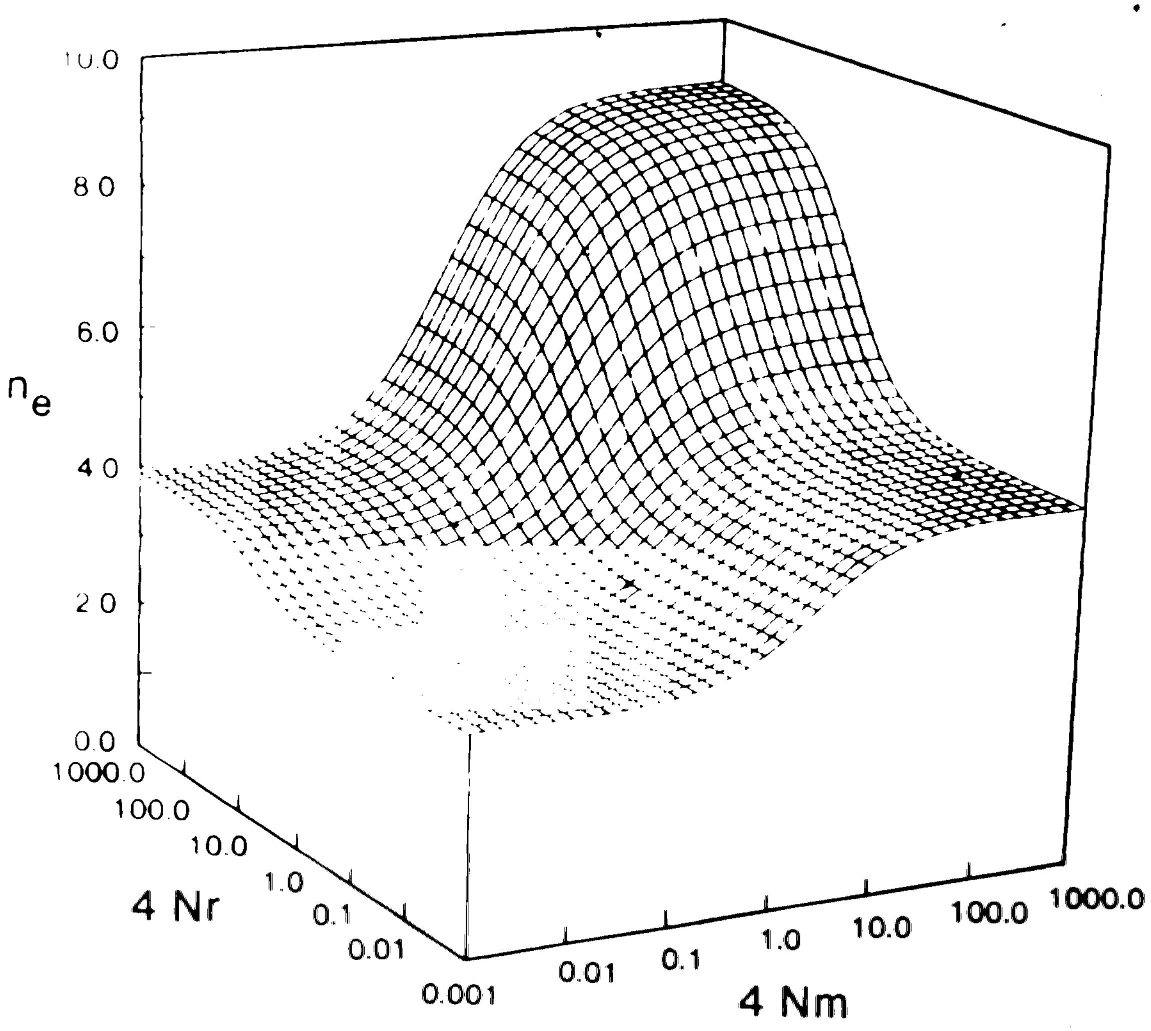
$$n_e^* = \frac{1}{\frac{1}{2}\Phi_{(ab)_1(ab)_1} + \frac{1}{2}\Phi_{(ab)_1(ab)_2}}$$

In a hybrid population between two races or species the genes would be sampled at random from each group and therefore $n_e^*$ should be used as the measure of variability.

The effective number of alleles at equilibrium in a single subpopulation, $n_e$, is given in Figure 4.1 for $4N\mu = 2.0$, $10^{-3} \leq 4Nr \leq 10^{3}$ and $10^{-3} \leq 4Nm \leq 10^{3}$. It shows that migration (hybridization) and recombination each significantly increases the number of alleles maintained in a population when $4Nm > 1$ or $4Nr > 1$. When both migration and recombination occur together the effective number of alleles is increased beyond the sum of the increases due to each process alone. This is because recombination requires initial variability to be present before it can generate more variability. The migration introduces new alleles which can then recombine with other alleles. It can be seen that the change in $n_e$ between low and high amounts of both migration and recombination is very large (an $n_e$ of 3 versus 9). Hence, the effective number of alleles in a natural population can be very large due to just these two processes, when both recombination and migration rates are sufficiently large. When the mutation rate is smaller than $4N\mu = 2.0$, the graphs show the same behavior but to a lesser degree. From studies on three sites it can be [illegible]

[illegible]

Figure 4.1: The equilibrium effective number of alleles maintained in each of the two subpopulations ($4N\mu$ = 2.0).

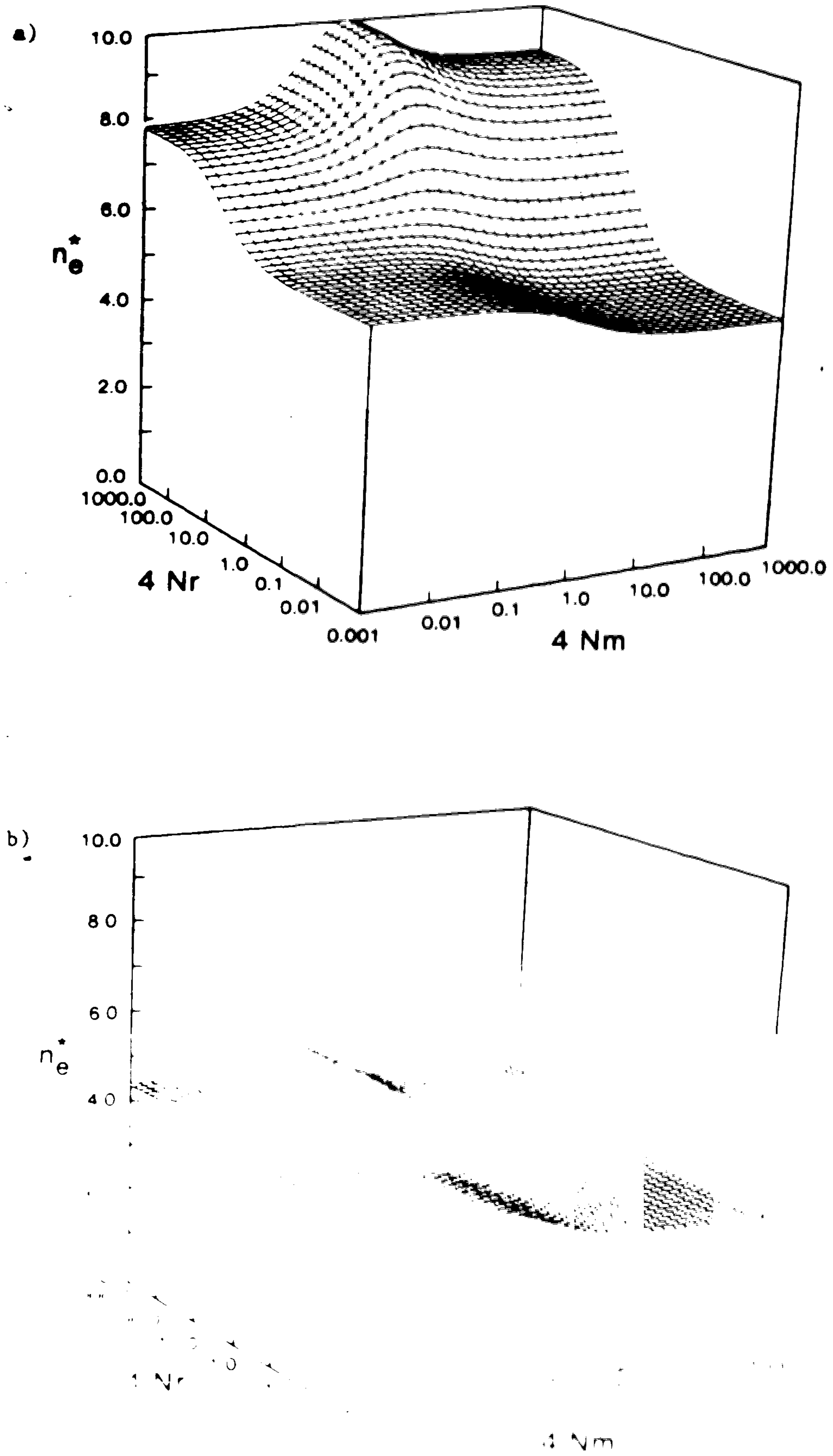

10.0
8 0
6.0
$n_e$
4 0
2 0
0.0
1000.0
100.0
10.0
1.0
0.1
0.01
0.001
4 Nr
0.01
0.1
1.0
10.0
100.0
1000.0
4 Nm

In a hybrid population the effective number of alleles at equilibrium is $n_e^*$. This quantity (plotted in Fig. 4.2 for $10^{-3} \le 4Nm \le 10^3$, $10^{-3} \le 4Nr \le 10^3$ and $4N\mu = 2.0$ (2a) or $4N\mu = 1.0$ (2b)) is quite different from the effective number of alleles in one subpopulation. Again, smaller mutation rates show the same relative effects as in Fig. 4.2b) but the effect is smaller in magnitude. There is an optimum when recombination is large and hybridization occurs at an intermediate rate. Since each subpopulation maintains a different array of alleles, hybridization can introduce new alleles to one population which can then recombine to create new variability. These recombinants are combinations which do not exist in either parental population and would appear as alleles unique to the hybrid population. Without recombination the variability introduced by hybridization is already present in the total population and does not increase $n_e^*$. Intragenic recombination can, therefore, account for the unique alleles seen in hybrid populations and perhaps for some of the mutants found in hybrid dysgenesis (Thompson and Woodruff, 1978).

The decrease on either side of the optimum as the hybridization rate changes is the result of several factors. First, new combinations will be created only if one of the a and b sites which recombine are of a form, or "allele", which does not exist in one of the subpopulations. Therefore, as the rate of hybridization increases and the similarity of alleles from each subpopulation increases the chance that recombinants will be new alleles decreases. On the other hand, a very small rate of hybridization does not introduce [illegible]

[illegible]

Figure 4.2: The equilibrium effective number of alleles maintained when gametes are sampled at random from each subpopulation (4.2a $4N\mu = 2.0$, 4.2b $4N\mu = 1.0$).


a)
10.0
8.0
6.0
$n_e^*$
4.0
2.0
0.0
1000.0
100.0
10.0
1.0
0.1
0.01
0.001
4 Nr
0.01
0.1
1.0
10.0
100.0
1000.0
4 Nm
b)
10.0
8 0
6 0
$n_e^*$
4 0
4 Nr
4 Nm

It has been shown by Malecot (1948) that the effective number of alleles in the entire population increases as the migration rates between the subpopulations decrease. As shown in Figure 4.2 this is not always true when intragenic recombination occurs. In a population with no recombination and no migration between subpopulations, $n_e^* = 2+8N\mu$, while in a population with large migration rates, $n_e^* = 1+8N\mu$. Therefore, for fixed $4N\mu$ and no recombination, a population with no migration between subpopulations always has more variability than with free migration. However, when the recombination value is high,

$$n_e^* = \frac{1}{\frac{1}{4}(\frac{1}{1+2N\mu})^2 + \frac{1}{2}(\ 0\ ) + \frac{1}{4}(\frac{1}{1+2N\mu})^2} = 2 + 8N\mu + 8N^2\mu^2$$

in a population with no migration between subpopulations. In a population with free migration

$$n_e^* = \frac{1}{\frac{1}{4}(\frac{1}{1+4N\mu})^2 + \frac{1}{2}(\frac{1}{1+4N\mu})^2 + \frac{1}{4}(\frac{1}{1+4N\mu})^2} = 1 + 8N\mu + 16N^2\mu^2$$

Thus, when the recombination value is high, a population with free migration between subpopulations will have more variability when 4N [illegible]. This difference is because the amount of variability intragenic recombination can create is a non linear, increasing

[illegible]

It has been suggested that some hybrid populations may be stable over long periods of time (Mayr, 1963, pp 368-379; Short, 1972; Hunt and Selander, 1973). In this case the equilibrium analysis above is appropriate. However, if the hybrid population is relatively recent it is necessary to consider the transient behavior. In order to investigate how the effective number of alleles changes over time in a non-equilibrium population, it is appropriate to assume that the two subpopulations initially are at equilibrium with no hybridization, ie: the equilibrium values of the coefficients when $m_1 = m_2 = 0$. Hybridization, at a constant rate, is then introduced and the change in the value of the coefficients over time is followed (again mutation rates, migration rates and population sizes are assumed to be equal for each of the subpopulations). Figure 4.3 shows the results with $n_e$ plotted for $N = 10^4$, $4N\mu = 2.0$, $4Nr = 0.1$, $1.0$, $10.0$ and $4Nm = 10.0$. The abscissa gives the number of generations starting at generation #0 (each subpopulation at equilibrium with no hybridization) and each generation up to #10,000. The equilibrium that will eventually be reached is indicated by an arrow. For all values of 4Nr, the transient value of $n_e$ (the effective number of alleles within one subpopulation) shows an increase above the eventual equilibrium as hybridization and recombination introduce new alleles to the subpopulation and thereafter a slow decline as random drift eliminates alleles. With increasing values of 4Nr the difference between the maximum $n_e$ and the eventual equilibrium becomes larger. In all cases the equilibrium is above the initial value and the transient $n_e$ becomes higher still.

Figure 4.4 gives the results for $n_e^*$ for the same parameter values

Figure 4.3: The effective number of alleles in each subpopulation after two isolated populations begin hybridization with each other ($4N\mu = 2.0$, $4Nm = 10.0$, $N = 10^4$).

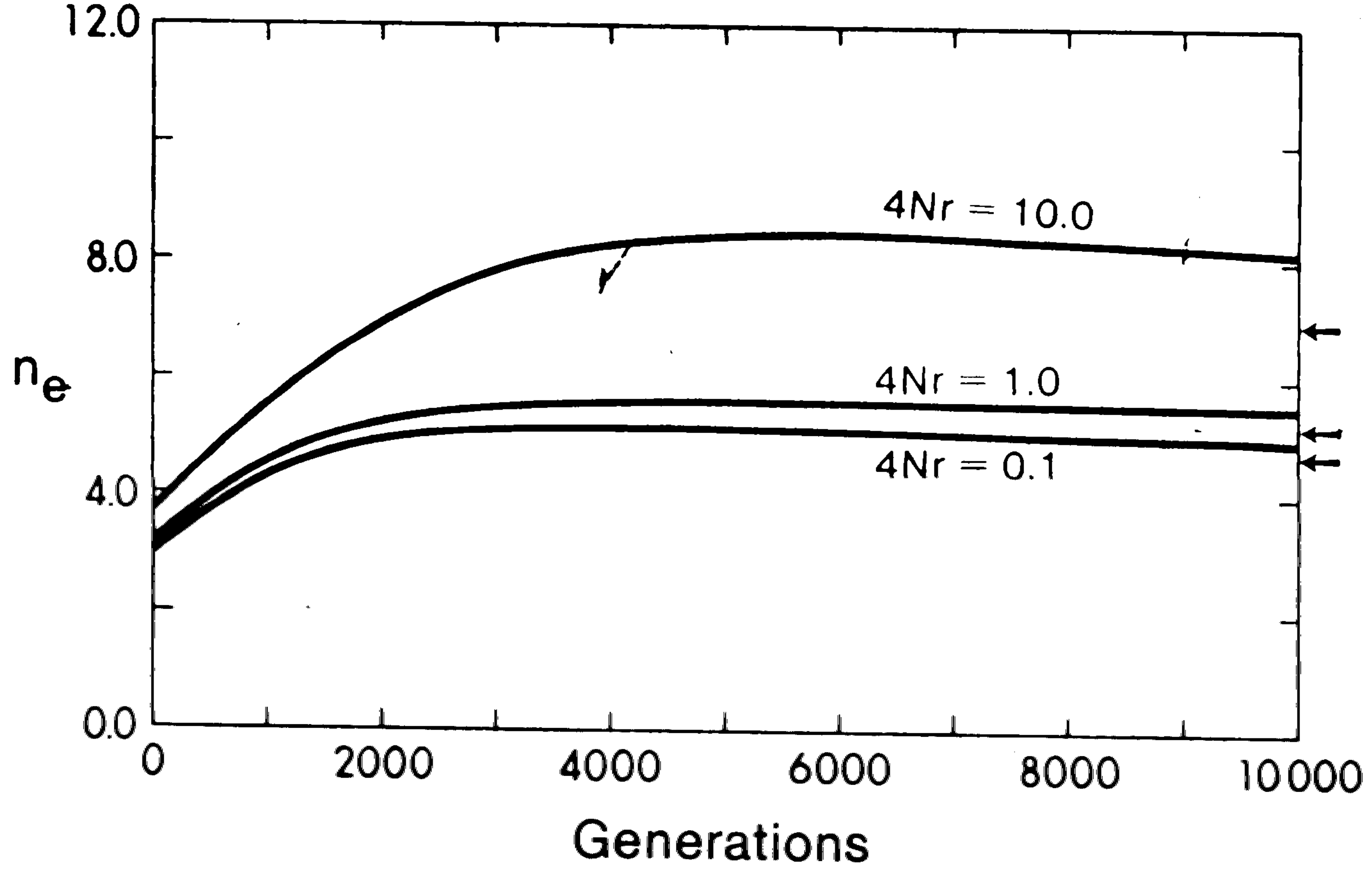
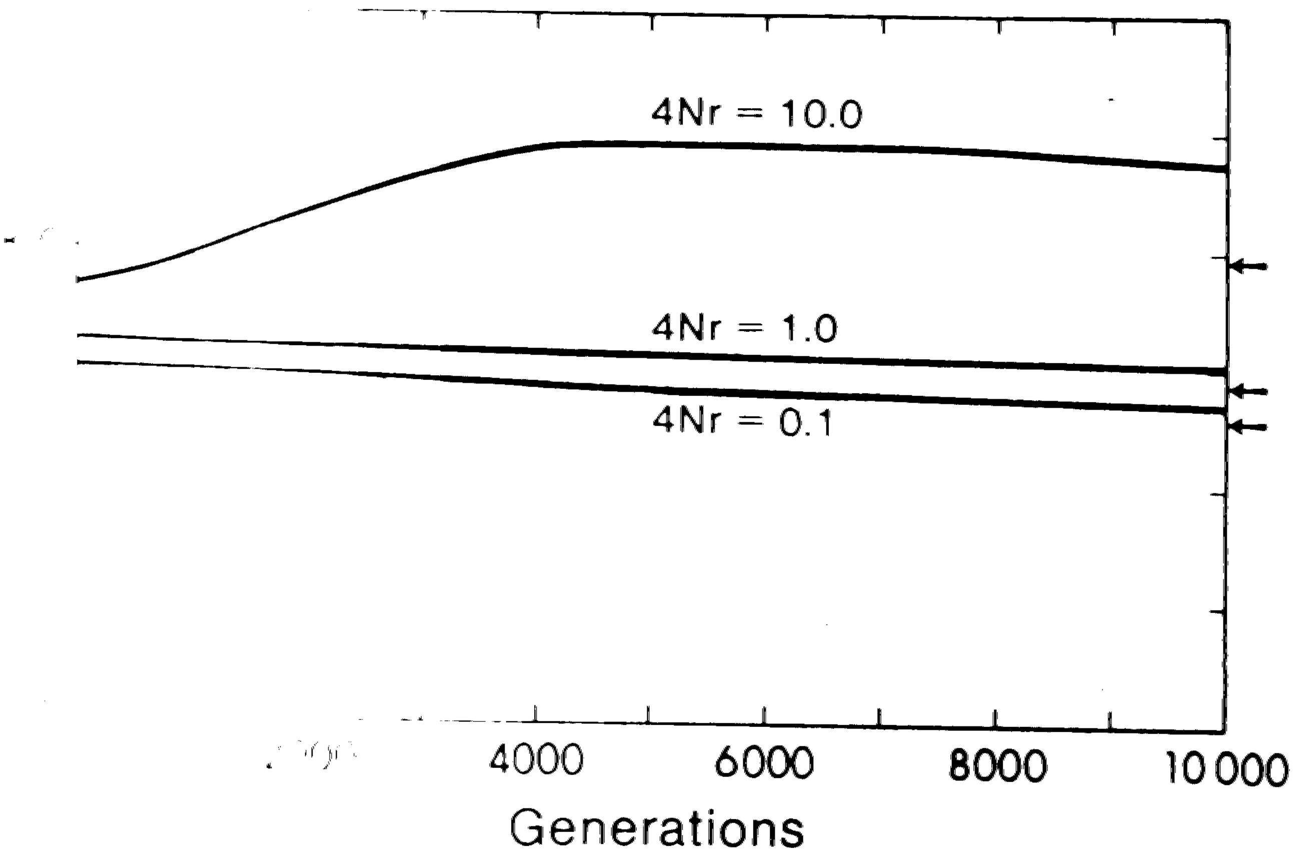

12.0
8.0
4.0
0.0
$n_e$
0
2000
4000
6000
8000
10000
Generations
4Nr = 10.0
4Nr = 1.0
4Nr = 0.1

Figure 4.4: The effective number of alleles maintained when gametes are sampled at random from each subpopulation after two isolated populations begin hybridization with each other (4Nu = 2.0, 4Nm = 10.0, N = [illegible]


4Nr = 10.0
4Nr = 1.0
4Nr = 0.1
4000
6000
8000
10 000
Generations

as Figure 4.3. For a small 4Nr, $n_e^*$ goes straight toward a lower equilibrium value (though very slowly); whereas with a large 4Nr, $n_e^*$ overshoots its equilibrium value which is larger than the initial value. Since $n_e^*$ is a measure of the total population variability, if the two subpopulations were to instantaneously mix there would be no change in $n_e^*$. Therefore, initially the only new variability in the total population is created by recombination. Without this recombination, new alleles would be formed only by mutation.

Figures 4.3 and 4.4 show that the effects of past events will be retained in a population for a very large number of generations. This makes the utility of an equilibrium analysis questionable, because the approach to such an equilibrium is very slow. For example, when 4Nm = 1.0 and 4Nr = 10.0, $n_e^*$ is still increasing away from the equilibrium even after 10,000 generations have passed. Similarly, $n_e$ must exceed the equilibrium before returning to it many generations later; however, when 4Nm = 1.0 and 4Nμ = 2.0, $n_e$ has not yet increased up to the equilibrium after 10,000 generations. Other hybridizations or events of importance to the population are certain to occur before an equilibrium is reached. Also the effects of hybridization can actually be larger in a recent hybrid population than those expected [illegible] an equilibrium population. In general $n_e$ increases above the equilibrium, though the length of time is prohibitive when 4Nm is small and $n_e^*$ also increases above the equilibrium when 4Nr is large. [illegible] may be difficult to observe the initial effects of hybridization on [illegible] parameters because $n_e^*$ changes slowly. From Figure 4.4 it [illegible] changes asymptotically. Indeed, a[illegible] [illegible] changes [illegible] is present [illegible] little differ[illegible] would be noticeable after 125 generations.

This study indicates that homozygosity is not an effective way to determine if intragenic recombination is an important factor in creating new alleles in a hybrid population. This determination might, however, be done at the molecular level by sequencing the DNA. If one part of the gene had a sequence characteristic of one subpopulation and another region of the gene was characteristic of the other subpopulation then intragenic recombination would be indicated.

## Summary

A two site, infinite allele model is used to study the influence of intragenic recombination on the effective number of neutral alleles in a hybrid population. It is shown that the combination of intragenic recombination and hybridization can have a large effect on the effective number of alleles in a population at equilibrium and an even larger effect when the population is not at equilibrium. When the mutation and recombination rates are large, a completely subdivided population will not maintain as much variability as a random mating population. It is concluded that unique alleles in hybrid populations could be formed by intragenic recombination.

# Chapter 5

# Variance and Covariance of Homozygosity in a Structured Population

## Introduction

The amount of homozygosity is a basic measure of variability in a natural population. The expected homozygosity in a finite population with selectively neutral alleles was first determined by Haldane (1939) and independently by Malecot (1948) and by Kimura and Crow (1964). To interpret observed levels of homozygosity it is also necessary to known the expected variance. The variance of homozygosity in a finite population with mutation and with selectively neutral alleles was determined by Watterson (1974) and by Stewart (1976), and the transient behavior by Li and Nei (1975).

A method to derive the variance of homozygosity is developed here, using identity coefficients. The method is applied to derive the variance and covariance of homozygosity for a structured population. The population is assumed to be divided into n partially isolated subpopulations each with N diploid, randomly mating individuals. The variance of homozygosity for completely isolated populations (including it's decomposition into component parts) has been derived by Lessard (1981). The variance of homozygosity within each subpopulation, the variance of homozygosity when gametes are sampled at random from the subpopulations and the covariance of homozygos[illegible]

[illegible]

those expected for a single random mating population.

## Variance of Homozygosity in a Single Population

The variance of homozygosity is first derived for a single population and the derivation is then extended to a structured population.

Consider a locus (denoted A) which can have k possible alleles. Let the population consist of N randomly mating diploid individuals. Each generation alleles can mutate to any particular allele at a rate $\mu\frac{1}{k-1}$ per gamete per generation. Thus the total mutation rate of an allele to any other allele is $\mu$ per gamete per generation.

When N >> 1, the expected "homozygosity" is

$$E(\sum_{i=1}^{k} p_i^2)$$

where $p_i$ is the frequency of the i-th allele and the variance of homozygosity is

$$Var(\sum_{i=1}^{k} p_i^2) = E[(\sum_{i=1}^{k} p_i^2)^2] - E(\sum_{i=1}^{k} p_i^2)^2$$

$$= E(\sum_{i=1}^{k} p_i^4) + E(\sum_{i=1}^{k}\sum_{\substack{j=1\\j\neq i}}^{k} p_i^2 p_j^2) - E(\sum_{i=1}^{k} p_i^2)^2$$

When N [illegible] and neglecting terms of order 1/N, the expected homozygosity is equal to the probability that two gametes, sampled at random without replacement, have the same allele at locus A (Kimura [illegible] [illegible]). This [illegible] can be denoted [illegible]. The terms in the formula for the variance of homozygosity can also be expressed as identity coefficients. The term $E(\sum_{i=1}^{k} p_i^2)$ is approximately the probability that four gametes sampled without replacement carry alleles at locus A which are identical in state and can be denoted $\Delta_{1111}$. The term $E(\sum_{i=1}^{k}\sum_{\substack{j=1 \\ j\neq i}}^{k} p_i^2 p_j^2)$ is approximately the probability that of four gametes sampled without replacement, two pairs have the same allele and each pair have different alleles. This probability can be denoted $\Delta_{11/11}$. Thus the variance of homozygosity can be expressed as

$$\mathrm{Var}(\sum_{i=1}^{k} p_i^2) = \Delta_{1111} + \Delta_{11/11} - \Phi_{11}^2$$

To determine recursion relationships for these coefficients it is convenient to define another six identity coefficients. Throughout, the coefficients are denoted by $\Phi$ if they involve a sample of two gametes, by $\Gamma$ if they involve a sample of three gametes and by $\Delta$ if they involve a sample of four gametes. A slash is used to separate non-identical genes; and subscripts denote the subpopulation from which the alleles were chosen. Denoting the allele on an arbitrary gamete by $a_i$, the nine coefficients can be defined as in Table 5.1.

Recursion relationships for the expected values of the coefficients over replicate populations are given in Appendix 4. Although these nine coefficients help to make the derivation more obvious, only four of them are necessary since

$$\Gamma_{111} + 3\Gamma_{11/1} + \Gamma_{1/1/1} = 1$$

$$\Delta_{1111} + 4\Delta_{111/1} + 3\Delta_{11/11} + 6\Delta_{11/1/1} + \Delta_{1/1/1/1} = 1$$

[illegible]

Table 5.1: Definitions of identity coefficients for a single locus in one population.

| | | |
|---|---|---|
| [illegible]$_{11}$ = Prob($a_1 \equiv a_2$) | $\Gamma_{111}$ = Prob($a_1 \equiv a_2 \equiv a_3$) | $\Gamma_{11/1}$ = Prob($a_1 \equiv a_2 \not\equiv a_3$) |
| [illegible]$_{1/1/1}$ = Prob($a_1 \not\equiv a_2 \not\equiv a_3$) | $\Delta_{1111}$ = Prob($a_1 \equiv a_2 \equiv a_3 \equiv a_4$) | $\Delta_{111/1}$ = Prob($a_1 \equiv a_2 \equiv a_3 \not\equiv a_4$) |
| [illegible] = Prob($a_1 \equiv a_2 \not\equiv a_3 \equiv a_4$) | $\Delta_{11/1/1}$ = Prob($a_1 \equiv a_2 \not\equiv a_3 \not\equiv a_4$) | $\Delta_{1/1/1/1}$ = Prob($a_1 \not\equiv a_2 \not\equiv a_3 \not\equiv a_4$) |

$$\Delta_{1111} + \Delta_{111/1} = \Gamma_{111}$$

$$\Delta_{111/1} + \Delta_{11/11} + \Delta_{11/1/1} = \Gamma_{11/1}$$

When $1/N$, $\mu << 1$ and using these identities, the recursion relationships reduce to those shown in Table 5.2. Therefore at equilibrium, with $\Lambda = 4N\mu\frac{1}{k-1}$,

$$\hat{\Phi}_{11} = \frac{1+\Lambda}{(1+k\Lambda)} \qquad \hat{\Gamma}_{111} = \frac{(1+\Lambda)(2+\Lambda)}{(1+k\Lambda)(2+k\Lambda)}$$

$$\hat{\Delta}_{1111} = \frac{(1+\Lambda)(2+\Lambda)(3+\Lambda)}{(1+k\Lambda)(2+k\Lambda)(3+k\Lambda)} \qquad \hat{\Delta}_{11/11} = \frac{\Lambda(1+\Lambda)^2(k-1)}{(1+k\Lambda)(2+k\Lambda)(3+k\Lambda)}$$

$$\hat{\mathrm{Var}}(\Sigma p_i^2) = \frac{2\Lambda(1+\Lambda)(k-1)}{(1+k\Lambda)^2(2+k\Lambda)(3+k\Lambda)}$$

This is the same result obtained by Watterson (1974) and Stewart (1976).

The variance has a maximum value, for particular values of $4N\mu$ and k, which results in several interesting properties. When k [illegible] a maximum variance of 0.0508 is obtained when

$$[illegible]$$

[illegible] the maximum variance is 0.0432 with [illegible]

maximum [illegible]

[illegible]

Table [illegible]. Approximate recursion relationships for the expected values of the identity coefficients over replicate populations. A single locus in one population.

$$\Phi'_{11} = \frac{1}{2N} + 2\mu\frac{1}{k-1} + (1- \frac{1}{2N} - 2\mu\frac{k}{k-1})\Phi_{11}$$

$$\Gamma'_{111} = (\frac{3}{2N} + 3\mu\frac{1}{k-1})\Phi_{11} + (1- \frac{3}{2N} - 3\mu\frac{k}{k-1})\Gamma_{111}$$

$$\Delta'_{1111} = (\frac{6}{2N} + 4\mu\frac{1}{k-1})\Gamma_{111} + (1- \frac{6}{2N} - 4\mu\frac{k}{k-1})\Delta_{1111}$$

$$\frac{1}{2N} + 4\mu\frac{1}{k-1})\Phi_{11} - (\frac{2}{2N} + 4\mu\frac{1}{k-1})\Gamma_{111} + (1- \frac{6}{2N} - 4\mu\frac{k}{k-1})\Delta_{11/11}$$

As shown by Li and Nei (1975) the system of equations can be solved at generation t and the variance is given by

$$\mathrm{Var}^{(t)}\Big(\sum_{i=1}^{k} p_i^2\Big) = \mathrm{Var}^{(\infty)}\Big(\sum_{i=1}^{k} p_i^2\Big) - \lambda_1^t(\alpha_1 - 2\phi_{11}^{(0)}\hat{\phi}_{11} + 2\hat{\phi}_{11}^2) +$$

$$\lambda_1^t(\phi_{11}^{(0)} - \hat{\phi}_{11})^2 + \lambda_2^t\alpha_2 + \lambda_3^t(\Delta_{1111}^{(0)} + \Delta_{11/11}^{(0)} - \hat{\Delta}_{1111} - \hat{\Delta}_{11/11} - \alpha_1 - \alpha_2)$$

where $\alpha_1 = 2(\phi_{11}^{(0)} - \hat{\phi}_{11})(16 + 10\Lambda + k\Lambda + k\Lambda^2)/(4 + k\Lambda)(5 + k\Lambda)$

$$\alpha_2 = 8(\Gamma_{111}^{(0)} - \hat{\Gamma}_{111} - 3(\phi_{11}^{(0)} - \hat{\phi}_{11})(2 + \Lambda)/(4 + k\Lambda))/(6 + k\Lambda)$$

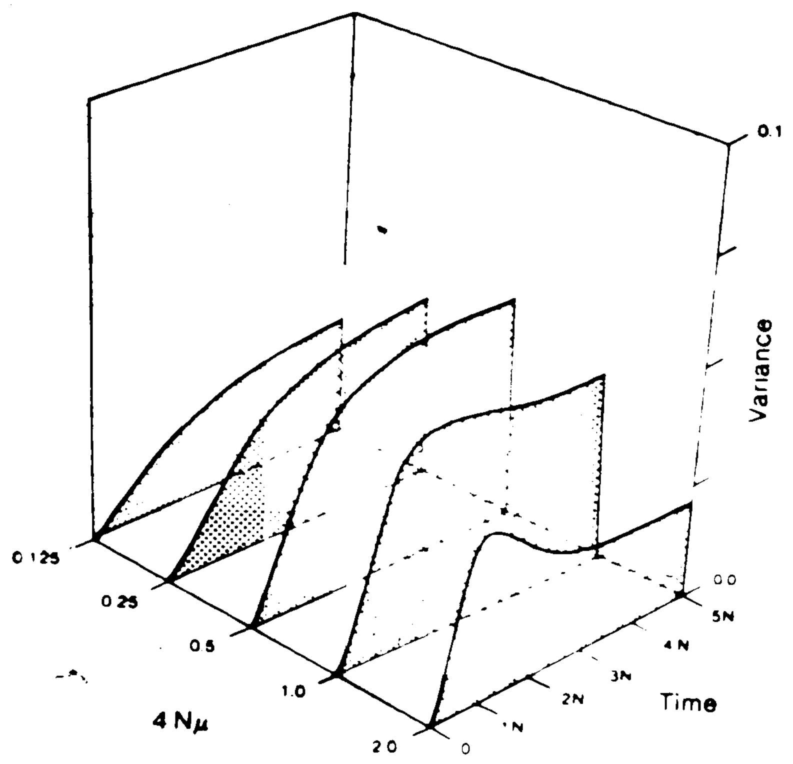
$$\lambda_1 = (1 - \frac{1}{2N} - 2\mu\frac{k}{k-1})$$

$$\lambda_2 = (1 - \frac{3}{2N} - 3\mu\frac{k}{k-1})$$

$$\lambda_3 = (1 - \frac{6}{2N} - 4\mu\frac{k}{k-1})$$

In Figure 5.1 the approach to equilibrium is shown, starting with a completely homozygous population. As time proceeds, variability increases within the population. When $4N\mu < 0.5$, the variance of homozygosity quickly increases and then asymptotically approaches its equilibrium value. When $4N\mu > 0.5$, the variance increases up to and then beyond its equilibrium value. This is because the variance is maximum when $4N\mu$ = 0.5. The increase in the variance is large when $4N\mu$ is large.

Figure 5.1: Variance of homozygosity in a single population over time, starting with a completely homozygous population.


0.1
Variance
0.0
0.125
0.25
0.5
1.0
2.0
4 Nμ
0
1N
2N
3N
4N
5N
Time

### Variance of homozygosity for a structured population

To determine the variance of homozygosity within subdivided populations consider n subpopulations each with N diploid individuals. Each generation a proportion m of the gametes in each subpopulation are migrants chosen at random from the remaining n-1 subpopulations. At equilibrium (or with initial conditions such that the probability of identity/non-identity is independent of the numeration of the subpopulations, eg: $\Delta_{iijk}$ for all $i \neq j \neq k$ are equal) this model requires a minimum of 17 coefficients. These coefficients are defined in Table 5.3, where $a_{ix}$ is an arbitrary gamete chosen from the i-th subpopulation. When 1/N, $\mu$, m << 1, recursion relationships for the expected values of the coefficients over replicate populations can be found and are given in Table 5.4. An analytical solution would be difficult to find for this system of equations. Therefore, the equations were solved numerically by substituting particular values for the subpopulation size, mutation rate and migration rate.

The variance of homozygosity within the i-th subpopulation is

$$\mathrm{Var}_i = \Delta_{iiii} + \Delta_{ii/ii} - \Phi_{ii}^2$$

and the covariance of homozygosity between the i-th and j-th subpopulations is

$$\mathrm{Cov}_{ij} = \Delta_{iijj} + \Delta_{ii/jj} - \Phi_{ii}^2$$

When gametes are sampled at random from the subpopulations the

Table 5.3: Definitions of identity coefficients for a single locus in a structured population.

$$\Phi_{ii} = \mathrm{Prob}(a_{i1} \equiv a_{i2}) \qquad \Phi_{ij} = \mathrm{Prob}(a_{i1} \equiv a_{j2}) \qquad \Gamma_{iii} = \mathrm{Prob}(a_{i1} \equiv a_{i2} \equiv a_{i3})$$

$$\Gamma_{iij} = \mathrm{Prob}(a_{i1} \equiv a_{i2} \equiv a_{j3}) \qquad \Gamma_{ijk} = \mathrm{Prob}(a_{i1} \equiv a_{j2} \equiv a_{k3}) \qquad \Delta_{iiii} = \mathrm{Prob}(a_{i1} \equiv a_{i2} \equiv a_{i3} \equiv a_{i4})$$

$$\Delta_{iiij} = \mathrm{Prob}(a_{i1} \equiv a_{i2} \equiv a_{i3} \equiv a_{j4}) \qquad \Delta_{iijj} = \mathrm{Prob}(a_{i1} \equiv a_{i2} \equiv a_{j3} \equiv a_{j4}) \qquad \Delta_{iijk} = \mathrm{Prob}(a_{i1} \equiv a_{i2} \equiv a_{j3} \equiv a_{k4})$$

$$\Delta_{ijk\ell} = \mathrm{Prob}(a_{i1} \equiv a_{j2} \equiv a_{k3} \equiv a_{\ell 4}) \qquad \Delta_{ii/ii} = \mathrm{Prob}(a_{i1} \equiv a_{i2} \not\equiv a_{i3} \equiv a_{i4}) \qquad \Delta_{ii/ij} = \mathrm{Prob}(a_{i1} \equiv a_{i2} \not\equiv a_{i3} \equiv a_{j4})$$

$$\Delta_{ij/ij} = \mathrm{Prob}(a_{i1} \equiv a_{j2} \not\equiv a_{i3} \equiv a_{j4}) \qquad \Delta_{ii/jj} = \mathrm{Prob}(a_{i1} \equiv a_{i2} \not\equiv a_{j3} \equiv a_{j4}) \qquad \Delta_{ij/ik} = \mathrm{Prob}(a_{i1} \equiv a_{j2} \not\equiv a_{i3} \equiv a_{k4})$$

$$\Delta_{ii/jk} = \mathrm{Prob}(a_{i1} \equiv a_{i2} \not\equiv a_{j3} \equiv a_{k4}) \qquad \Delta_{ij/k\ell} = \mathrm{Prob}(a_{i1} \equiv a_{j2} \not\equiv a_{k3} \equiv a_{\ell 4})$$

where $i \neq j \neq k \neq \ell$

[illegible] relationships for the expected values of the identity [illegible] populations. A single locus in a structured population.

$$\phi'_{ii} = \frac{1}{2N} + 2\nu\frac{1}{k-1} + (1 - \frac{1}{2N} - 2\nu\frac{k}{k-1} - 2m)\phi_{ii} + 2m\phi_{ij}$$

$$\phi'_{ij} = 2\nu\frac{1}{k-1} + (1 - 2\nu\frac{k}{k-1} - 2m\frac{1}{n-1})\phi_{ij} + (2m\frac{1}{n-1})\phi_{ii}$$

$$\Gamma'_{iii} = (\frac{3}{2N} + 3\nu\frac{1}{k-1})\phi_{ii} + (1 - \frac{3}{2N} - 3\nu\frac{k}{k-1} - 3m)\Gamma_{iii} + 3m\Gamma_{iij}$$

$$\Gamma'_{iij} = (\nu\frac{1}{k-1})\phi_{ii} + (\frac{1}{2N} + 2\nu\frac{1}{k-1})\phi_{ij} + (1 - \frac{1}{2N} - 3\nu\frac{k}{k-1} - m\frac{2n-3}{n-1})\Gamma_{iij} + (m\frac{1}{n-1})\Gamma_{iii} + (2m\frac{n-2}{n-1})\Gamma_{ijk}$$

$$\Gamma'_{ijk} = (3\nu\frac{1}{k-1})\phi_{ij} + (1 - 3\nu\frac{k}{k-1} - 3m\frac{2}{n-1})\Gamma_{ijk} + (3m\frac{2}{n-1})\Gamma_{iij}$$

$$\Delta'_{iiii} = (\frac{6}{2N} + 4\nu\frac{1}{k-1})\Gamma_{iii} + (1 - \frac{6}{2N} - 4\nu\frac{k}{k-1} - 4m)\Delta_{iiii} + 4m\Delta_{iiij}$$

$$\Delta'_{iiij} = (\nu\frac{1}{k-1})\Gamma_{iii} + (\frac{3}{2N} + 3\nu\frac{1}{k-1})\Gamma_{iij} + (1 - \frac{3}{2N} - 4\nu\frac{k}{k-1} - m\frac{3n-2}{n-1})\Delta_{iiij} + (m\frac{1}{n-1})\Delta_{iiii} + (3m\frac{1}{n-1})\Delta_{iijj} + (3m\frac{n-2}{n-1})\Delta_{iijk}$$

$$\Delta'_{iijj} = (\frac{2}{2N} + 4\nu\frac{1}{k-1})\Gamma_{iij} + (1 - \frac{2}{2N} - 4\nu\frac{k}{k-1} - 4m)\Delta_{iijj} + (4m\frac{1}{n-1})\Delta_{iiij} + (4m\frac{n-2}{n-1})\Delta_{iijk}$$

$$\Delta'_{iijk} = (2\nu\frac{1}{k-1})\Gamma_{iij} + (\frac{1}{2N} + 2\nu\frac{1}{k-1})\Gamma_{ijk} + (1 - \frac{1}{2N} - 4\nu\frac{k}{k-1} - 2m)\Delta_{iijk} + (2m\frac{1}{n-1})\Delta_{iiij} + (2m\frac{1}{n-1})\Delta_{iijj} + (2m\frac{n-3}{n-1})\Delta_{ijk\ell}$$

$$\Delta'_{ijk\ell} = (4\nu\frac{1}{k-1})\Gamma_{ijk} + (1 - 4\nu\frac{k}{k-1} - 12m\frac{1}{n-1})\Delta_{ijk\ell} + (12m\frac{1}{n-1})\Delta_{iijk}$$

$$\Delta'_{ii/ii} = (\frac{2}{2N} + 4\mu\frac{1}{k-1})\Phi_{ii} - (\frac{2}{2N} + 4\mu\frac{1}{k-1})\Gamma_{iii} + (1 - \frac{6}{2N} - 4\mu\frac{k}{k-1} - 4m)\Delta_{ii/ii} + 4m\Delta_{ii/ij}$$

$$\Delta'_{ii/ij} = (2\mu\frac{1}{k-1})\Phi_{ii} + (\frac{1}{2N} + 2\mu\frac{1}{k-1})\Phi_{ij} - (\mu\frac{1}{k-1})\Gamma_{iii} - (\frac{1}{2N} + 3\mu\frac{1}{k-1})\Gamma_{iij} + (1 - \frac{3}{2N} - 4\mu\frac{k}{k-1} - m\frac{3n-2}{n-1})\Delta_{ii/ij} + (m\frac{1}{n-1})\Delta_{ii/ii} +$$
$$(2m\frac{1}{n-1})\Delta_{ij/ij} + (m\frac{1}{n-1})\Delta_{ii/jj} + (m\frac{n-2}{n-1})\Delta_{ii/jk} + (2m\frac{n-2}{n-1})\Delta_{ij/ik}$$

$$\Delta'_{ij/ij} = (4\mu\frac{1}{k-1})\Phi_{ij} - (4\mu\frac{1}{k-1})\Gamma_{iij} + (1 - \frac{2}{2N} - 4\mu\frac{k}{k-1} - 4m)\Delta_{ij/ij} + (4m\frac{1}{n-1})\Delta_{ii/ij} + (4m\frac{n-2}{n-1})\Delta_{ij/ik}$$

$$\Delta'_{ii/jj} = (\frac{2}{2N} + 4\mu\frac{1}{k-1})\Phi_{ii} - (\frac{2}{2N} + 4\mu\frac{1}{k-1})\Gamma_{iij} + (1 - \frac{2}{2N} - 4\mu\frac{k}{k-1} - 4m)\Delta_{ii/jj} + (4m\frac{1}{n-1})\Delta_{ii/ij} + (4m\frac{n-2}{n-1})\Delta_{ii/jk}$$

$$\Delta'_{ii/jk} = (2\mu\frac{1}{k-1})\Phi_{ii} + (\frac{1}{2N} + 2\mu\frac{1}{k-1})\Phi_{ij} - (2\mu\frac{1}{k-1})\Gamma_{iij} - (\frac{1}{2N} + 2\mu\frac{1}{k-1})\Gamma_{ijk} + (1 - \frac{1}{2N} - 4\mu\frac{k}{k-1} - 2m\frac{n+1}{n-1})\Delta_{ii/jk} + (2m\frac{1}{n-1})\Delta_{ii/ij} +$$
$$(2m\frac{1}{n-1})\Delta_{ii/jj} + (4m\frac{1}{n-1})\Delta_{ij/ik} + (2m\frac{n-3}{n-1})\Delta_{ij/k\ell}$$

$$\Delta'_{ij/ik} = (4\mu\frac{1}{k-1})\Phi_{ij} - (2\mu\frac{1}{k-1})\Gamma_{iij} - (2\mu\frac{1}{k-1})\Gamma_{ijk} + (1 - \frac{1}{2N} - 4\mu\frac{k}{k-1} - 2m\frac{n}{n-1})\Delta_{ij/ik} + (2m\frac{1}{n-1})\Delta_{ii/ij} + (2m\frac{1}{n-1})\Delta_{ij/ij} +$$
$$(2m\frac{1}{n-1})\Delta_{ii/jk} + (2m\frac{n-3}{n-1})\Delta_{ij/k\ell}$$

$$\Delta'_{ij/k\ell} = (4\mu\frac{1}{k-1})\Phi_{ij} - (4\mu\frac{1}{k-1})\Gamma_{ijk} + (1 - 4\mu\frac{k}{k-1} - 12m\frac{1}{n-1})\Delta_{ij/k\ell} + (4m\frac{1}{n-1})\Delta_{ii/jk} + (8m\frac{1}{n-1})\Delta_{ij/ik}$$

apparent homozygosity is approximately

$$\frac{1}{n}\Phi_{ii} + \frac{n-1}{n}\Phi_{ij}$$

and the variance of homozygosity is approximately

$$\text{Var} = \frac{1}{n^3}\Big(\Delta_{iiii} + 4(n-1)\Delta_{iiij} + 3(n-1)\Delta_{iijj} + 6(n-1)(n-2)\Delta_{iijk} +$$

$$(n-1)(n-2)(n-3)\Delta_{ijk\ell} + \Delta_{ii/ii} + 4(n-1)\Delta_{ii/ij} + 2(n-1)\Delta_{ii/ij} +$$

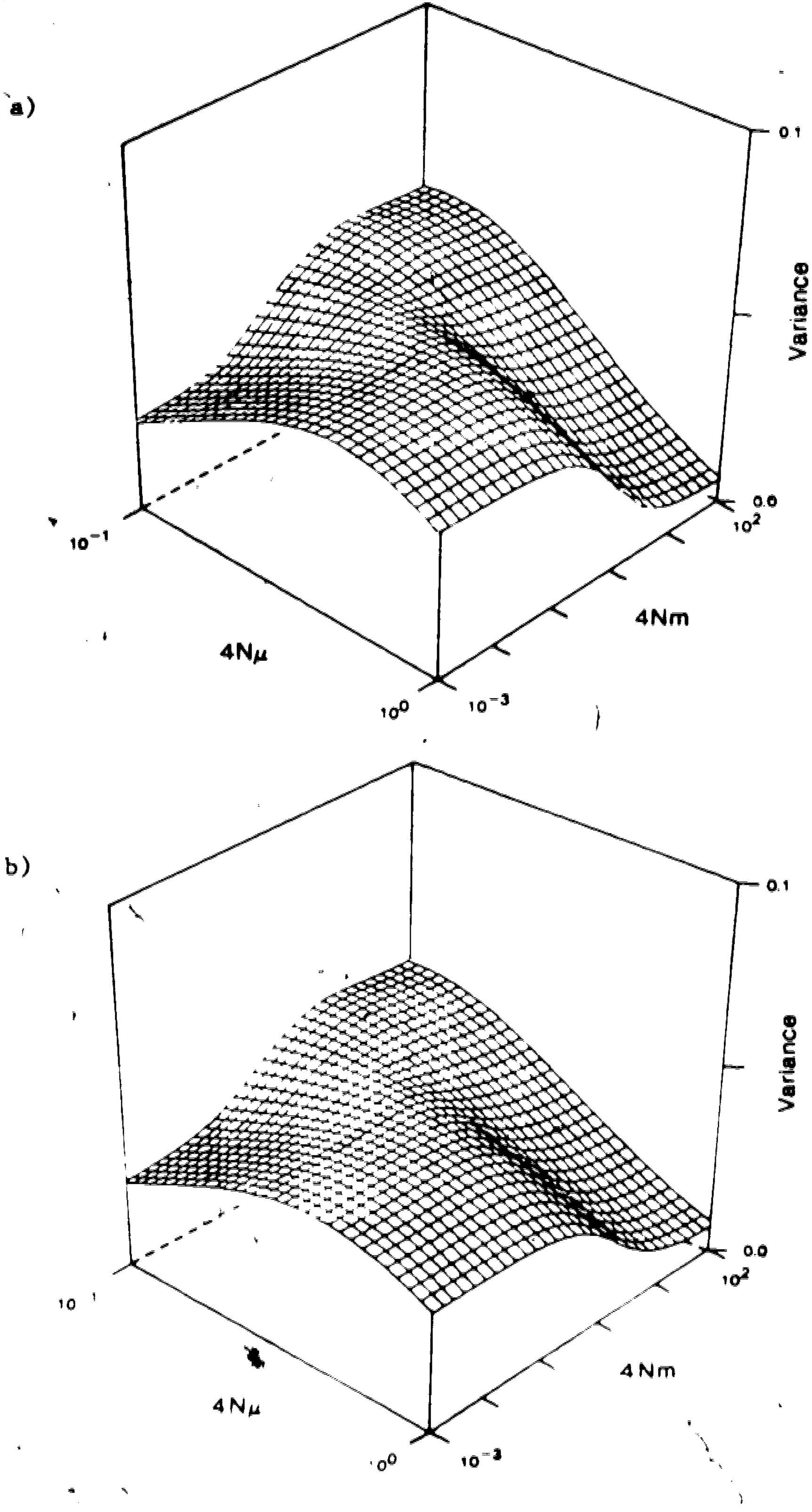
$$(n-1)\Delta_{ii/jj} + 2(n-1)(n-2)\Delta_{ii/jk} + 4(n-1)(n-2)\Delta_{ij/ik} +$$

$$(n-1)(n-2)(n-3)\Delta_{ij/k\ell}\Big) - \left(\frac{1}{n}\Phi_{ii} + \frac{n-1}{n}\Phi_{ij}\right)^2 .$$

The following results are for k → ∞ or k = 4 and n = 4. The results for other values of k and n are similar unless stated otherwise.

The variance of homozygosity within one of four subpopulations for k → ∞ and for k = 4 is given in Figure 5.2 with $10^{-1} \leq 4N\mu \leq 10^{5}$ and $10^{-3} \leq 4Nm \leq 10^{2}$. The results illustrate the effect of the maximum variance when $4N\mu \simeq 0.5$. Depending on the amount of migration the equilibrium variance will be higher or lower than that expected in a single population. This is because migrants can carry new alleles with them, augmenting the mutation rate so that it is closer to or exceeds the maximum. When k = 4 (as is appropriate for a single nucleotide), the results remain qualitatively the same. When migration occurs between a larger number of subpopulations, the migrants are more likely to carry different alleles and thus the variance changes faster as the migration rate changes.

Figure 5.2: Equilibrium variance of homozygosity within one of four subpopulations (5.2a k → ∞, 5.2b k = 4).


a)
0.1
Variance
0.0
$10^2$
4Nm
$10^{-3}$
$10^0$
4Nμ
$10^{-1}$
b)
0.1
Variance
0.0
$10^2$
4Nm
$10^{-3}$
$10^0$
4Nμ
$10^{-1}$

When the subdivision is unknown to an observer or a hybrid population is considered, the expected variance of homozygosity is given by sampling gametes at random from each subpopulation. The result of this is shown in Figure 5.3 with a) k → ∞, b) k = 4 and with $10^{-1} \leq 4N\mu \leq 10^{0}$ and $10^{-3} \leq 4Nm \leq 10^{2}$. When the migration rate is small, the results change as k changes. This is because the probability of picking identical alleles from two subpopulations is zero when k → ∞ and m = 0, but this probability is 1/4 when k = 4. Thus, the variance of homozygosity is small when k → ∞ but remains relatively large when k = 4. This effect is more dramatic when n < 4.

To determine the transient behavior of the variance of homozygosity, each subpopulation is assumed to be initially at equilibrium with m = 0. Migration, at a constant rate, is then introduced and the change in the value of the coefficients over time is followed by iterating the equations in Table 5.4. This is done for 4Nm = 10.0 and k → ∞ and the results are shown in Figure 5.4. The equilibrium that will eventually be reached is indicated by an arrow. Within a single subpopulation (Figure 5.4a) there is a large and rapid decrease in the variance and then a slow increase back to equilibrium. Lower migration rates cause slower rates of change and higher migration rates cause faster rates of change, but the results remain similar. Note that the time scale is very large in these graphs. For example, if $N = 10^{4}$ and a generation length of 20 years for man is assumed, the time scale covers more than half a million years. Nevertheless, the value of the variance of homozygosity at time t = 0 (when $4N\mu = 0.125$) is closer to its equilibrium value than at time t = 3N. When gametes are chosen at random from the subpopulations, the

Figure 5.3: Equilibrium variance of homozygosity sampling at random from four subpopulations (5.3a k → ∞, 5.3b k = 4).

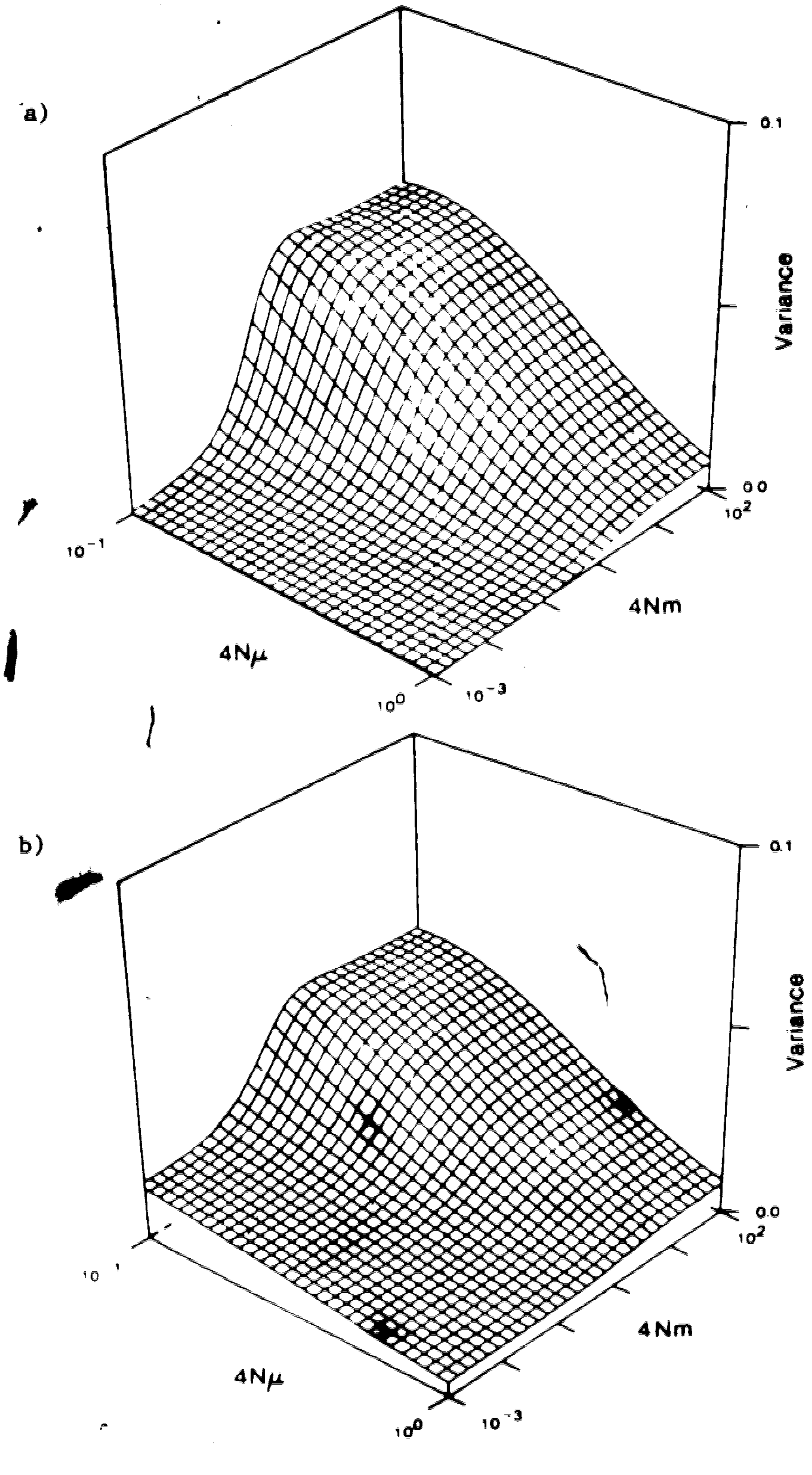
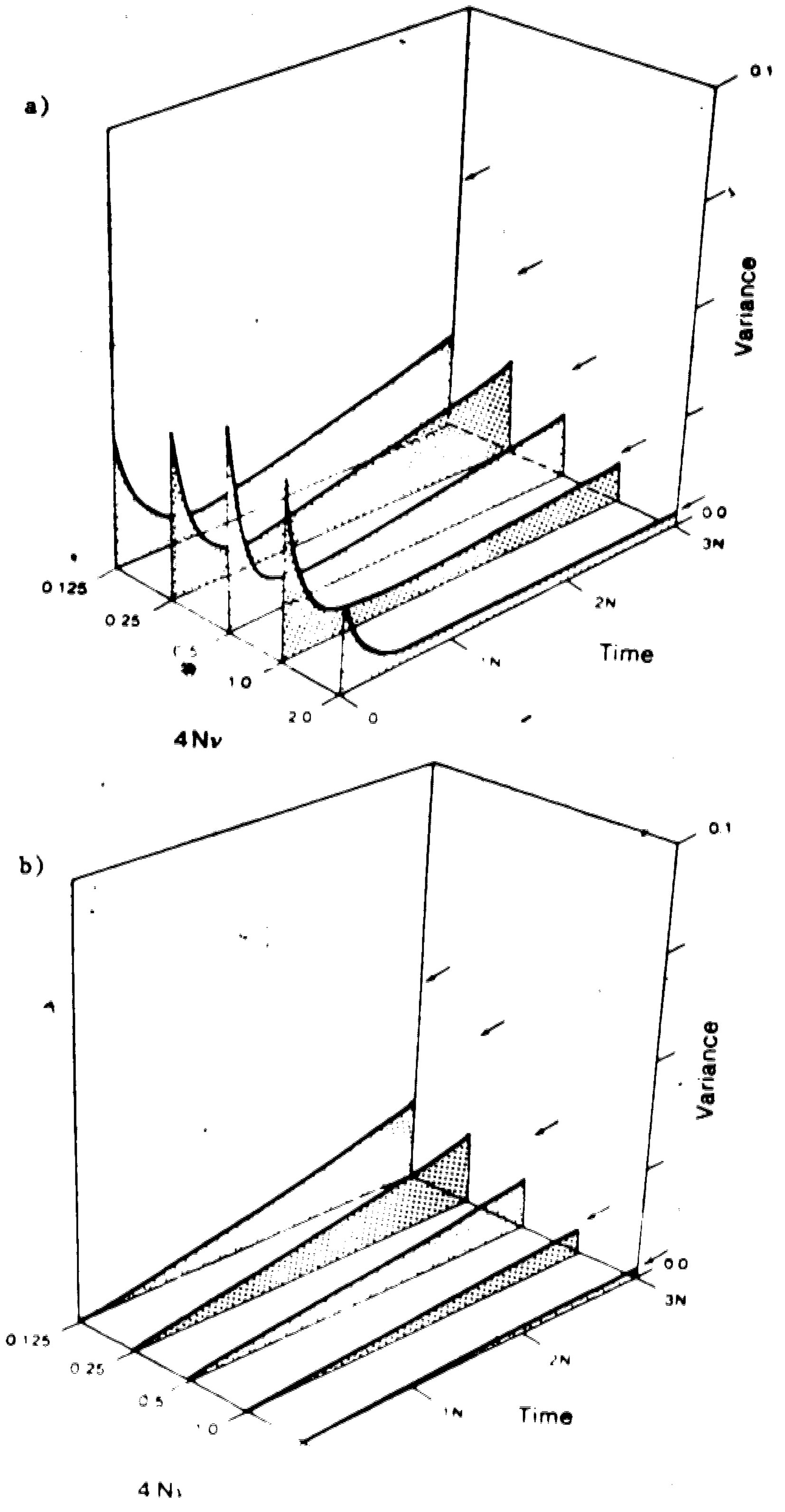

a)
0.1
Variance
0.0
$10^2$
4Nm
$10^{-3}$
$10^0$
4Nμ
$10^{-1}$
b)
0.1
Variance
0.0
$10^2$
4Nm
$10^{-3}$
$10^0$
4Nμ
$10^{-1}$

Figure 5.4: Transient variance of homozygosity within one of four subpopulations (5.4a $k \to \infty$) and the transient variance of homozygosity sampling at random from four subpopulations (5.4b, $k \to \infty$).

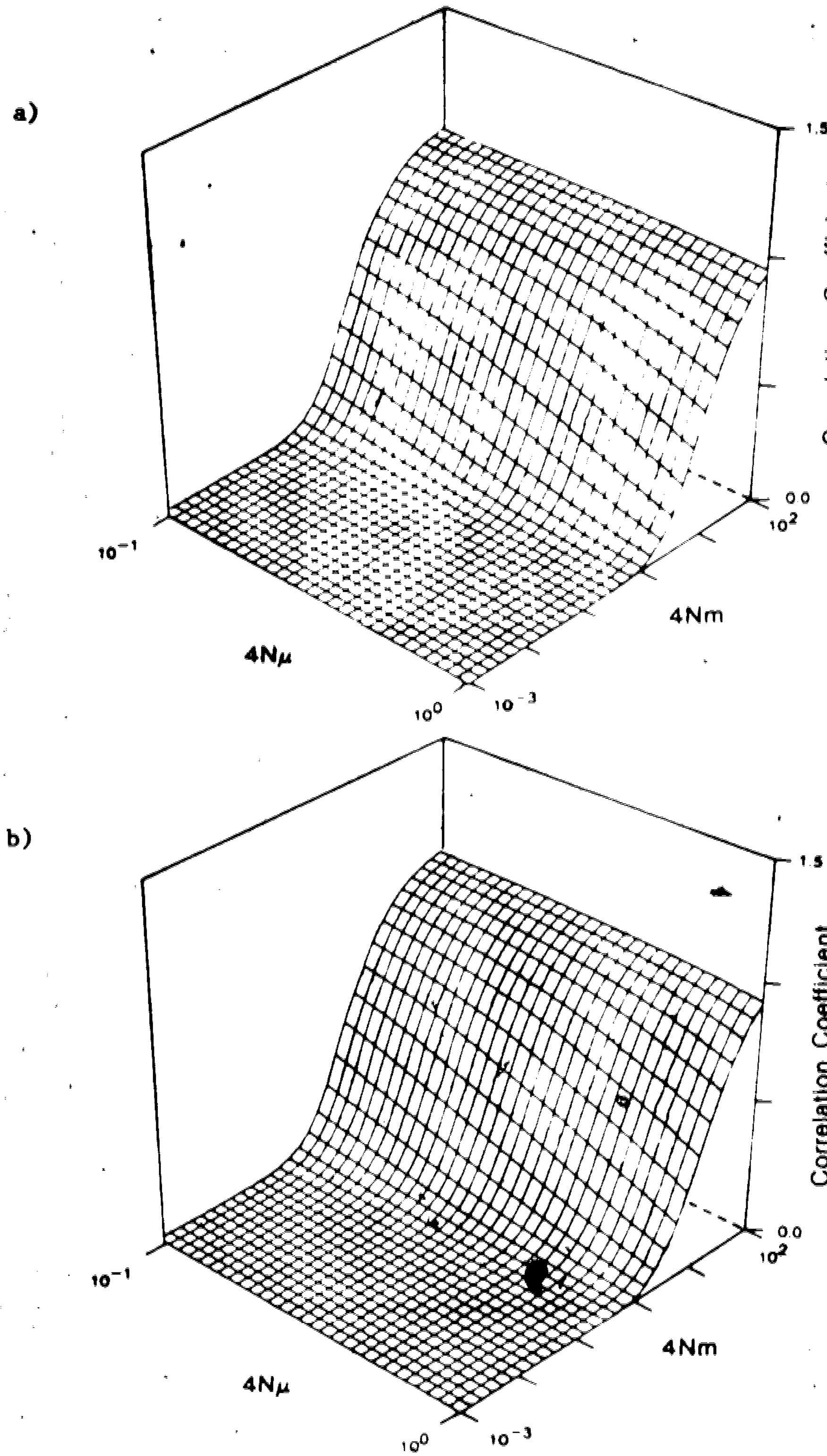

a)
0.1
Variance
0.0
3N
2N
1N
Time
0
0.125
0.25
0.5
1.0
2.0
4Nν
b)
0.1
Variance
0.0
3N
2N
1N
Time
0.125
0.25
0.5
1.0
4Nν

results shown in Figure 5.4b are obtained. In this case there is not a drastic decrease in the variance. In general there is a slow, monotonic increase in the variance toward the new equilibrium. Again when $4N\mu$ is small it takes a large number of generations to approach the equilibrium. When k and n are small, the variance of homozygosity, picking gametes at random from the subpopulations, does not decrease over time. In this case the initial variance of homozygosity is close to the equilibrium value and only minor changes occur. The variance of homozygosity within a subpopulation also shows smaller changes when k is small.

Figure 5.5 shows the correlation coefficient for homozygosity between two subpopulations with $10^{-1} \leq 4N\mu \leq 10^{0}$ and $10^{-3} \leq 4Nm \leq 10^{2}$. The correlation coefficient is almost identical whether $k \to \infty$ (5.5a) or $k = 4$ (5.5b). Malécot (1948) showed that when $4Nm > 1$, the alleles of two subpopulations are very similar ($\Phi_{ij} \approx \Phi_{ii}$). Figure 5.5 shows that this is also true for the second moment of frequencies for several subpopulations and with an infinite or finite number of alleles. Over a short range of 4Nm, the subpopulations change from being unrelated to strongly correlated. Presumably this is also true for all gene frequency moments.

Most natural populations are subdivided in some way and this creates problems for many statistical tests. For example Ewens' (1972) method for estimating $4N_{\mu}$ considers only a single population. To extend the method for a structured population would be difficult. It is, therefore, necessary to known how strongly the subdivision affects the variance of homozygosity (and higher moments) relative to that

Figure 5.5: Equilibrium correlation coefficient of homozygosity between two of four subpopulations (5.5a $k \to \infty$, 5.5b $k = 4$).


a)
1.5
Correlation Coefficient
0.0
$10^2$
$10^{-1}$
4Nm
4Nμ
$10^0$
$10^{-3}$
b)
1.5
Correlation Coefficient
0.0
$10^2$
$10^{-1}$
4Nm
4Nμ
$10^0$
$10^{-3}$

expected in a single population. To do this we have used the homozygosity within a subpopulation to calculate the appropriate value of $4N\mu$ for a single population, from the relation

$$4N\mu = \frac{1 - \Phi_{ii}}{\frac{k}{k-1}\Phi_{ii} - \frac{1}{k-1}}$$

This value of $4N\mu$ was then substituted into Stewart's (1976) formula to give an expected variance of homozygosity. The ratio of the true variance to this expected value is given in Figure 5.6. As can be seen, this ratio is close to one for all $4N\mu$, $4Nm$ and $k \to \infty$, $k = 4$. The maximum and minimum of the ratio is 1.039 and 0.881 in Figure 5.6a) and 1.051 and 0.986 in Figure 5.6b), respectively. When $n = 2$, the ratio is even closer to one. If the behavior of higher order moments are reflected by that of the variance, this suggests that many statistical tests may be appropriately applied to a subpopulation which has migration with other subpopulations. Similar results were found in the simulations of Ewens and Gillespie (1974) and Slatkin (1982).

If however the subdivision is unknown to an observer, this is no longer true. Figure 5.7 shows the ratio of the true variance to the expected variance when genes are sampled at random from four subpopulations. When k is large, the actual variance of homozygosity is much smaller than the variance appropriate for the expected level of homozygosity. Therefore, an observer must know the subdivisions of the population under study. The simulation studies of Ewens and Gillespie (1974) suggested that population subdivision, with larger migration rates, does not invalidate the use of Ewens' theory. The

Figure 5.6: Ratio of the actual variance of homozygosity within one of four subpopulations to the expected variance for a single population with the same variability (5.6a $k \to \infty$, 5.6b $k = 4$).

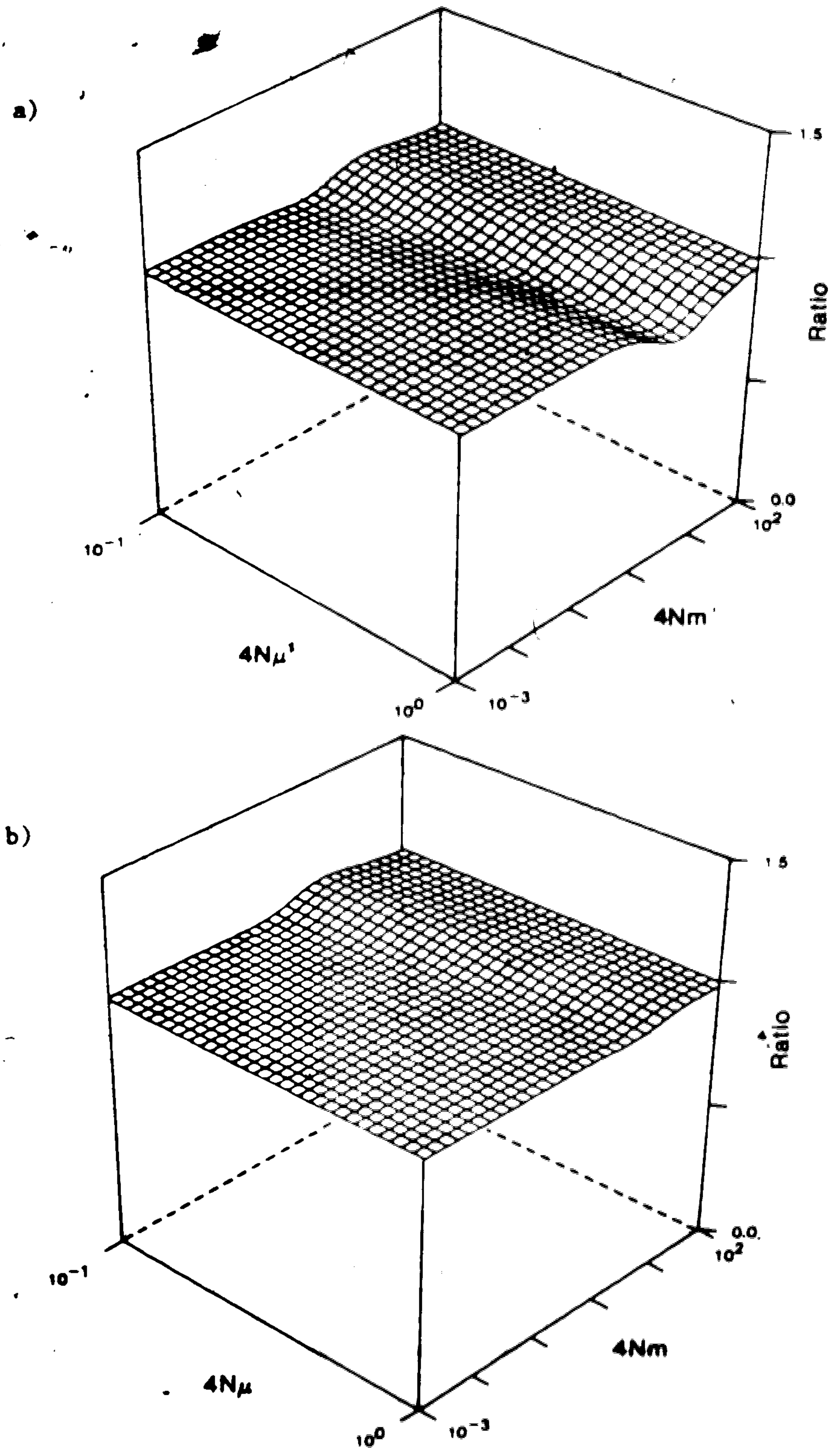
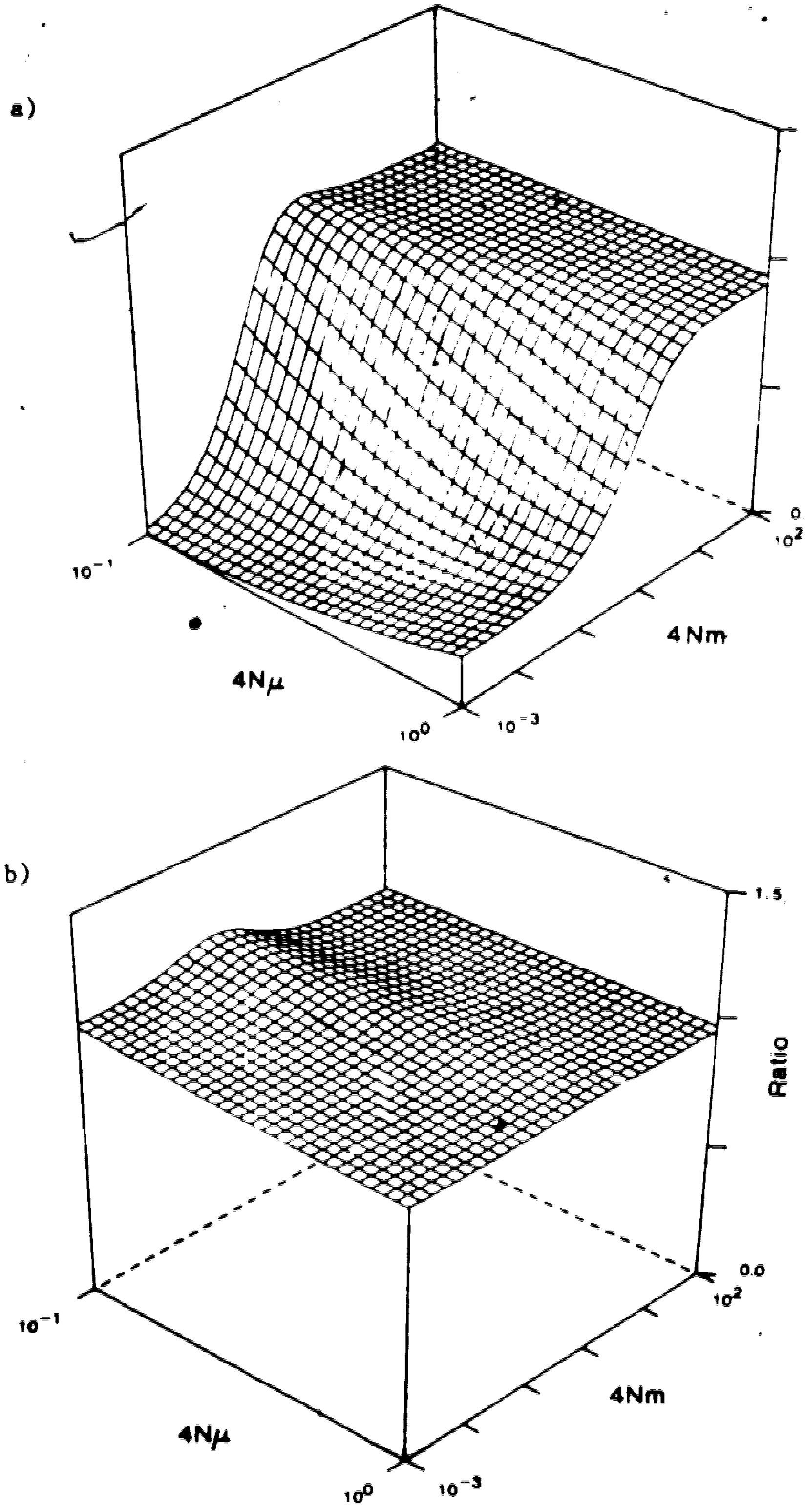

a)
1.5
Ratio
0.0
10^2
4Nm
10^-3
10^0
4Nμ'
10^-1
b)
1.5
Ratio
0.0
10^2
4Nm
10^-3
10^0
4Nμ
10^-1

Figure 5.7: Ratio of the actual variance of homozygosity sampling gametes at random from four subpopulations to the expected variance for a single population with the same variability (5.7a $k \to \infty$, 5.7b $k = 4$).


a)
1.5
Ratio
0.0
10^2
4Nm
10^-3
10^0
4Nμ
10^-1
b)
1.5
Ratio
0.0
10^2
4Nm
10^-3
10^0
4Nμ
10^-1

simulation of Slatkin (1982) suggests the opposite when the migration rates are small. As noted by Slatkin this is due to the difference in the migration rates. From Figure 5.7a) we can quantify the range of [illegible] where this change takes place as $10^{\text{[illegible]}} < 4Nm < 10^{1}$. Figure 5.7b) shows that the ratio is again close to one when k is small. The maximum and minimum of the ratio is 1.057 and 0.061 in Figure 5.7a) and 1.139 and 0.995 in Figure 5.7b), respectively. Thus even if a population is subdivided, estimates of the mutation rates of nucleotides (eg: using Ewens' 1974 method) may be appropriate.

## Summary

The variance of homozygosity for a k-allele model with n partially isolated subpopulations is derived numerically using identity coefficients. Within a single population the variance has a maximum of approximately 0.05. Thus the transient variance may increase and then decrease over time when $4N\mu > 0.5$. This maximum also causes the variance within a subpopulation to depend strongly upon the migration rates with other subpopulations. The variance is not strongly influenced by the number of alleles possible at a locus unless the population is presumed panmictic, but is actually subdivided. When the latter is true, the variance is higher with small migration rates when k is small. The transient behavior of the variance of homozygosity shows that a large number of generations may be required to approach equilibrium values. The results suggest that, in many situations, the variance of homozygosity may be adequately estimated from the amount of variability present. If the results for higher order moments are similar, statistical tests need not consider structured populations as a special case.

# Chapter 6

# Two-Locus, Fourth Order Gene Frequency Moments: Implications for the Variance of Squared Linkage Disequilibrium and the Variance of Homozygosity

## Introduction

The partial differential equations for a diffusion approximation which describe the behavior of two, linked, neutral loci in a finite population have been known for a long time (Kimura, 1955). However, the equations are too complicated to be easily solved. One way to circumvent this problem is to evaluate only the moments of gene frequencies. It is known (eg: Hill and Robertson, 1968; Serant and Villard, 1972; Weir and Cockerham 1974; Serant, 1974; Hill, 1975; Li and Nei, 1975; Strobeck and Morgan, 1978; and others) that these moments follow simple recursion relationships which can be solved. This approach is used here to evaluate the two-locus, fourth order gene frequency moments. The results are applied to two problems; the variance of the expected squared linkage disequilibrium and the variance of homozygosity of a gene with recombination between two sites.

Linkage disequilibrium is a measure of nonrandom association between alleles at different loci. It can be defined as

$$D_{ij} = f_{ij} - p_i q_j$$

where $f_{ij}$ is the frequency of gametes with the i-th allele at locus A and j-th allele at locus B and where $p_i$ and $q_j$ are the corresponding allele frequencies. Both selection and random drift due to a finite population size can cause a nonrandom association. The sum of squares of the linkage disequilibria,

$$\sum_{ij}\sum D_{ij}^2 = \sum_{ij}\sum (f_{ij} - p_i q_j)^2$$

is a measure of the average disequilibria between two loci. It is a component of the squared correlation of gene frequencies, enters into the standard Chi square test and is of interest itself. Therefore, it is necessary to know the size of the variance that can be expected in a natural population. The results presented here demonstrate that the standard deviation of $(\sum_{ij}\sum D_{ij}^2)$ will usually be lar than the mean and can be much larger for realistic mutation rates.

The method can also be used to find the variance of homozygosity for a gene consisting of two sites between which recombination occurs. It is known that recombination occurs within genes and the discovery that introns are prevalent in most eukaryotic genes facilitates such recombination. It has been shown by monte-carlo simulation (Strobeck and Morgan, 1978) that intragenic recombination significantly increases the variance of homozygosity if $4N\mu > 1.0$ and $r > \mu$. The results confirm and extend this prediction.

## Theory

Consider two loci (denoted A and B) in a finite population with 2N gametes. The gametes in each generation are produced following a Wright-Fisher model (Ewens, 1979). Let the mutation rate to unique, selectively neutral alleles be $\nu_1$ and $\nu_2$ per gamete per generation, at loci A and B, respectively. Let the recombination rate between the two loci be r. Throughout, it is assumed that $1/2N$, $\nu_1$, $\nu_2$, $r << 1$ and terms of order $(1/2N)^2$, $\nu_1^2$, $\nu_2^2$, $r^2$ or higher are neglected.

To define the necessary system of equations to find fourth order moments requires a minimum of 50 identity coefficients, each the probability that a particular sample of gametes have (or do not have) identical alleles. These coefficients can be denoted

$$\Phi_{ijk/\ell mn/p/q}$$

To define these coefficients, consider a sample of $i+j+k+\ell+m+n+p+q$ gametes drawn at random, without replacement (a group of i gametes, a group of j gametes and so on). Signify the gametes in each of these groups with superscript roman numerals, eg: let $a_x^I$ (or $b_x^I$) denote the allele at locus A (or locus B) from the x-th gamete of the first group of i gametes and $a_x^{II}$ denote the allele at locus A from the x-th gamete of the second group of j gametes, etc. Define

$$\Phi_{ijk/\ell mn/p/q} = \text{Prob}[a_1^I \equiv \cdots \equiv a_i^I \equiv a_1^{II} \equiv \cdots \equiv a_j^{II} \equiv a_1^{VII} \equiv \cdots \equiv a_p^{VII}$$

$$\not\equiv a_1^{IV} \equiv \cdots \equiv a_\ell^{IV} \equiv a_1^{V} \equiv \cdots \equiv a_m^{V} \equiv a_1^{VIII} \equiv \cdots \equiv a_q^{VIII}$$

$$\text{and } b_1^I \equiv \cdots \equiv b_i^I \equiv b_1^{III} \equiv \cdots \equiv b_k^{III} \equiv b_1^{VIII} \equiv \cdots \equiv b_q^{VIII}$$

$$\not\equiv b_1^{IV}\equiv\cdots\equiv b_\ell^{IV}\equiv b_1^{VI}\equiv\cdots\equiv b_n^{VI}\equiv b_1^{VII}\equiv\cdots\equiv b_p^{VII}]$$

where "≡" should be read "is identical to" and "≢" should be read "are not identical to". These coefficients are defined diagrammatically Figure 6.1 and in in words, $\ddagger_{ijk/\ell mn/p/q}$ is the probability that (i) all genes at locus A carried by the $i+j+p$ gametes are identical; (ii) all genes at locus A carried by the $\ell+m+q$ gametes are identical; (iii) that the alleles (i) and (ii) are different; (iv) all genes at locus B carried by the $i+k+q$ gametes are identical; (v) all genes at locus B carried by the $\ell+n+p$ gametes are identical; and (vi) that the alleles in (iv) and (v) differ. Note that these coefficients apply to a specific sample of gametes. For example, the probability that any one of three genes differ is $3\Phi_{020/010/0/0}$ while $\ddagger_{020/010/0/0}$ is the probability that a particular gene differs. In general (with i, j, k, ℓ, m, n, p and q gametes in each group), the number of ways gametes can be sampled is

$$\frac{(i+j+k+\ell+m+n+p+q)!}{i!j!k!\ell!m!n!p!q!x}$$

where, if $i=\ell=p=q=0$ then

$x=4$ if $j=m\neq0$ and $k=n\neq0$

$x=2$ if $j=m\neq0$ or $k=n\neq0$

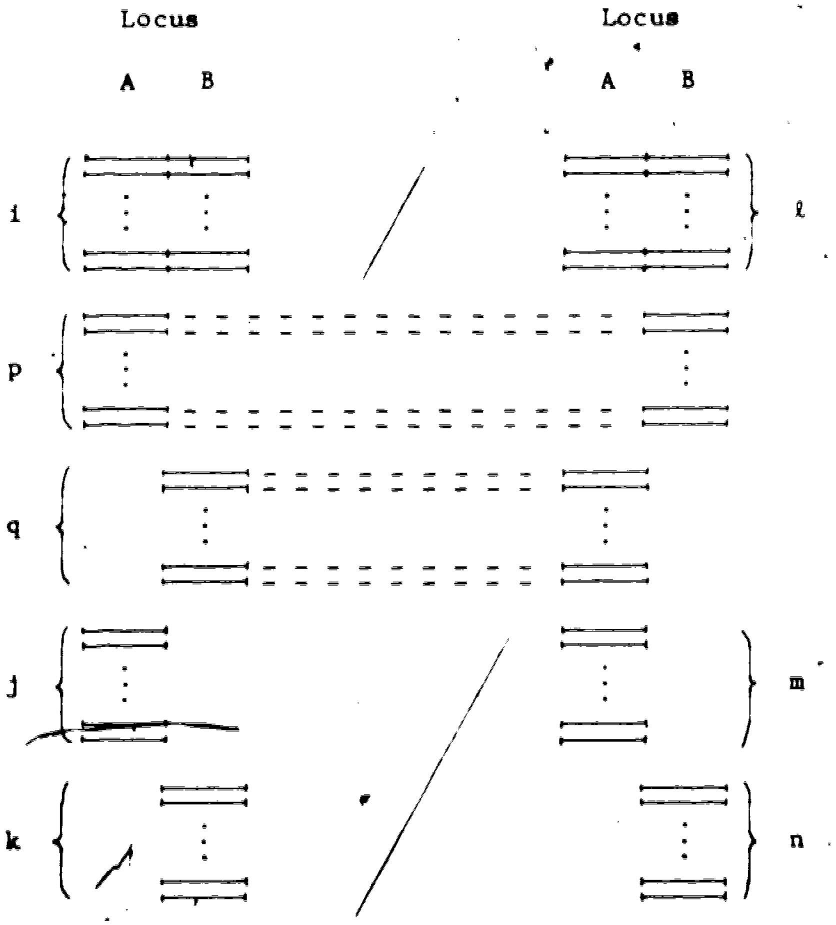
$x=1$ otherwise

if i, ℓ, p or $q\neq0$ then

$x=4$ if $i=\ell=p=q$ and $j=m$ and $k=n$

$x=2$ if $i=q$ and $\ell=p$ and $j=m$

Figure 6.1: Diagrammatic definition of the identity coefficients (All of the alleles in a vertical column must be identical, a slash separates non-identical alleles and the letters indicate the number of gametes sampled).


Locus
A
B
Locus
A
B
i
p
q
j
k
ℓ
m
n

x=2 if i=p and ℓ=q and k=n

x=2 if i=ℓ and p=q and j=m and k=n

x=1 otherwise

For convenience, let

$$\Phi_{ijk/000/0/0} = \Phi_{ijk}$$

The general recursion relationship for the expected values of the coefficients over replicate populations is derived in Appendix 5. The set of equations, necessary to determine the variance of linkage disequilibrium, is given in Appendix 6. Particular values for the recombination and mutation rate were substituted into these equations. The equations were then numerically solved on a computer to determine the equilibrium values of the identity coefficients.

The variance of squared linkage disequilibrium can be expressed as

$$\begin{aligned}
\mathrm{Var}[\sum_{ij}\sum D^2_{ij}] &= E[(\sum_{ij}\sum D^2_{ij})^2] - E[\sum_{ij}\sum D^2_{ij}]^2 \\
&= E[\sum\sum\sum_{ijk\ell}\sum(f^2_{ij}f^2_{k\ell}-2f^2_{ij}f_{k\ell}p_kq_\ell+f^2_{ij}p^2_kq^2_\ell-2f_{ij}f^2_{k\ell}p_iq_j+4f_{ij}f_{k\ell}p_ip_kq_jq_\ell \\
&\quad -2f_{ij}p_ip^2_kq_jq^2_\ell+f^2_{k\ell}p^2_iq^2_j-2f_{k\ell}p^2_ip_kq^2_jq_\ell+p^2_ip^2_kq^2_jq^2_\ell)] \\
&\quad - E[\sum_{ij}\sum(f^2_{ij}-2f_{ij}p_iq_j+p^2_iq^2_j)]^2 \\
&= E[\sum_{ij}\sum(f^4_{ij}-4f^3_{ij}p_iq_j+6f^2_{ij}p^2_iq^2_j-4f_{ij}p^3_iq^3_j+p^4_iq^4_j)] \\
&\quad + E[\sum\sum_{ijk\neq i}\sum(f^2_{ij}f^2_{kj}-2f^2_{ij}f_{kj}p_kq_j+f^2_{ij}p^2_kq^2_j-2f_{ij}f^2_{kj}p_iq_j+4f_{ij}f_{kj}p_ip_kq^2_j \\
&\quad -2f_{ij}p_ip^2_kq^3_j+f^2_{kj}p^2_iq^2_j-2f_{kj}p^2_ip_kq^3_j+p^2_ip^2_kq^4_j)] \\
&\quad + E[\sum\sum_{ij\ell\neq j}\sum(f^2_{ij}f^2_{i\ell}-2f^2_{ij}f_{i\ell}p_iq_\ell+f^2_{ij}p^2_iq^2_\ell-2f_{ij}f^2_{i\ell}p_iq_j+4f_{ij}f_{i\ell}p^2_iq_jq_\ell \\
&\quad -2f_{ij}p^3_iq_jq^2_\ell+f^2_{i\ell}p^2_iq^2_j-2f_{i\ell}p^3_iq^2_jq_\ell+p^4_iq^2_jq^2_\ell)]
\end{aligned}$$

$$+\ E[\sum_{i}\sum_{j}\sum_{k\neq i}\sum_{\ell\neq j}(f_{ij}^2f_{k\ell}^2-2f_{ij}^2f_{k\ell}p_kq_\ell+f_{ij}^2p_k^2q_\ell^2-2f_{ij}f_{k\ell}^2p_iq_j$$

$$+4f_{ij}f_{k\ell}p_ip_kq_jq_\ell-2f_{ij}p_ip_k^2q_jq_\ell^2+f_{k\ell}^2p_i^2q_j^2-2f_{k\ell}p_i^2p_kq_j^2q_\ell+p_i^2p_k^2q_j^2q_\ell^2)]$$

$$-\ E[\sum_{i}\sum_{j}(f_{ij}^2-2f_{ij}p_iq_j+p_i^2q_j^2)]^2$$

where the sums on i and k extend over all alleles at locus A and j and ℓ over all alleles at locus B. Each of these gene frequency moments is equivalent to an identity coefficient. In particular,

$$\Phi_{ijk/\ell mn/p/q}=E\left(\sum_{a_1}\sum_{b_1}\sum_{a_2\neq a_1}\sum_{b_2\neq b_1}f^{i}_{a_1b_1}f^{\ell}_{a_2b_2}f^{p}_{a_1b_2}f^{q}_{a_2b_1}p^{j}_{a_1}p^{m}_{a_2}q^{k}_{b_1}q^{n}_{b_2}\right)$$

Therefore, the variance can be expressed as

$$\mathrm{Var}(\sum_{i}\sum_{j}D_{ij}^2)=\Phi_{400}-4\Phi_{311}+6\Phi_{222}-4\Phi_{133}+\Phi_{044}$$

$$+\Phi_{200/000/0/2}-4\Phi_{201/010/0/1}+2\Phi_{202/020/0/0}+4\Phi_{112/010/0/1}$$

$$-4\Phi_{113/020/0/0}+\Phi_{024/020/0/0}+\Phi_{200/000/2/0}-4\Phi_{210/001/1/0}$$

$$+2\Phi_{220/002/0/0}+4\Phi_{121/001/1/0}-4\Phi_{131/002/0/0}+\Phi_{042/002/0/0}$$

$$+\Phi_{200/200/0/0}-4\Phi_{200/111/0/0}+2\Phi_{200/022/0/0}+4\Phi_{111/111/0/0}$$

$$-4\Phi_{111/022/0/0}+\Phi_{022/022/0/0}-(\Phi_{200}-2\Phi_{111}+\Phi_{022})^2$$

Similarly, for a gene with two sites the variance of homozygosity is

$$\mathrm{Var}\left(\sum_{i}\sum_{j}f_{ij}^2\right)=E\left((\sum_{i}\sum_{j}f_{ij}^2)^2\right)-E^2\left(\sum_{i}\sum_{j}f_{ij}^2\right)$$

$$=E\left(\sum_{i}\sum_{j}f_{ij}^4+\sum_{i}\sum_{j}\sum_{k\neq i}f_{ij}^2f_{kj}^2+\sum_{i}\sum_{j}\sum_{\ell\neq j}f_{ij}^2f_{i\ell}^2+\sum_{i}\sum_{j}\sum_{k\neq i}\sum_{\ell\neq j}f_{ij}^2f_{k\ell}^2\right)-E^2\left(\sum_{i}\sum_{j}f_{ij}^2\right)$$

and in terms of identity coefficients it is

$$=\Phi_{400}+\Phi_{200/000/0/2}+\Phi_{200/000/2/0}+\Phi_{200/200/0/0}-\Phi_{200}^2$$

## Results & Discussion

### Variance of squared linkage disequilibrium

Significant levels of linkage disequilibrium are found frequently in partially and completely selfing populations and between alleles associated with inversions. In other cases, extensive surveys of natural populations (Lewontin, 1974; Nevo, 1978; Brown, 1979) generally show only low levels of linkage disequilibrium. However, the variance of the expected squared linkage disequilibrium is not known.

Hill (1977) attempted to determine the coefficient of variation (C.V.) for the squared correlation coefficient

$$r^2 = E[D^2/p(1-p)q(1-q)]$$

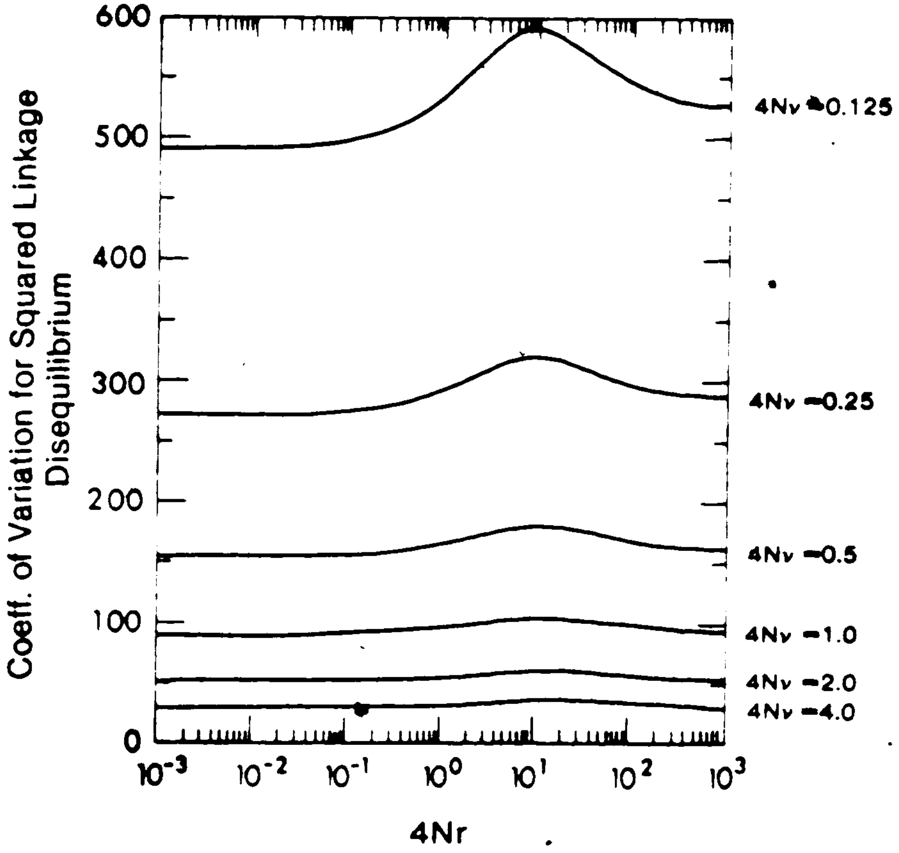
(where $D = f_{11} - p\,q$) for a two allele model in segregating populations. His results, using a Taylor's series approximation for both the mean and variance, indicated that the C.V. could be greater than one hundred percent. However, the remainder term in the Taylor's series for the mean can be large. This can result in a negative approximation to the squared correlation coefficient if there are rare alleles in the population. This is shown with an example given in Appendix 7. This example assumes that 12 replicate populations are observed, ten with the most frequent gamete having a frequency of 0.9998, and the other two replicate populations with the most frequent gamete having a frequency of 0.97 (ten populations are given rare alleles in order to make the effects of the rare alleles more noticeable). When the expected values for these populations are substituted into the second order Taylor's expansion of $r^2$

$$r^2 = \frac{E[D^2]}{E[p(1-p)q(1-q)]}\left(1 - \frac{E[D^2p(1-p)q(1-q)]}{E[D^2]E[p(1-p)q(1-q)]} + \frac{E[p^2(1-p)^2q^2(1-q)^2]}{E^2[p(1-p)q(1-q)]}\right)$$

a negative squared correlation is found. Since the Taylor's approximation to the squared correlation coefficient can be negative when the mutation rate is small, the C.V. for the correlation was not calculated.

The coefficient of variation for linkage disequilibrium is shown in Figure 6.2 with $\nu_1 = \nu_2 = \nu$, for $4N\nu$ = 0.125, 0.25, 0.5, 1.0, 2.0, 4.0. It can be seen that the C.V. will be less than one hundred percent only when $4N\nu$ is very large. When $4N\nu$ is small, the standard deviation is several times the size of the mean. The results in Figure 6.2 show that the C.V. is relatively constant for each $4N\nu$ when $4Nr$ is large or small. For small $4N\nu$, the minimum C.V. is reached as $r \to 0$ and the maximum when $4Nr \simeq 10.0$. Note also that as $r \to \infty$ the expected linkage disequilibrium approachs zero but the C.V. remains relatively constant. This demonstrates that the distribution must be highly skewed. The C.V. appears to increase exponentially as $4N\nu$ decreases. Since the C.V. is so large the utility of tests based on $D^2$ must be questioned.

Figure 6.2: The percent coefficient of variation for the squared linkage disequilibrium.


Coeff. of Variation for Squared Linkage Disequilibrium
600
500
400
300
200
100
0
10^-3
10^-2
10^-1
10^0
10^1
10^2
10^3
4Nr
4Nv =0.125
4Nv =0.25
4Nv =0.5
4Nv =1.0
4Nv =2.0
4Nv =4.0

Variance of homozygosity

The coefficient of variation of homozygosity when a gene consists of two sites with recombination between them can also be determined. This model is a good approximation of a gene with two exons and a single intron. Introns occur in most eukaryotic genes and must significantly increase recombination within genes since they can exist in large numbers and can be a major portion of the gene. For example; the vitellogenin genes of *Xenopus laevis* have 33 introns within each gene (Wahli *et al*., 1980); The α2 type I collagen gene of chickens has more than 49 introns (Vogeli *et al*., 1981); the single intron of the chloroplast $tRNA^{Ile}$ gene in *Zea mays* is more than 92% of the total length of the gene (Koch *et al*., 1981).

In Figure 6.3 and Figure 6.4 let $\mu = 2\nu$ be the mutation rate of the complete gene (each site within the gene is assumed to have the same mutation rate, $\nu$). Figure 6.3 shows that the C.V. of homozygosity increases when recombination occurs and when the mutation rate of the complete gene is large, as predicted by Strobeck and Morgan (1978). When the r to $\mu$ ratio is small, the maximum C.V. occurs when $\Theta = 8N\nu \approx 1.5$ (Figure 6.3). When the mutation rate is small, the effects of recombination become smaller since recombination acts only on variability already present. The variance of homozygosity when r is large is given by

$$\frac{4\Theta(6+6\Theta+\Theta^2)}{(1+\Theta)^4(2+\Theta)^2(3+\Theta)^2}$$

One possible way to adjust for the effects of intragenic

Figure 6.3: The percent coefficient of variation for the homozygosity of a gene consisting of two sites.

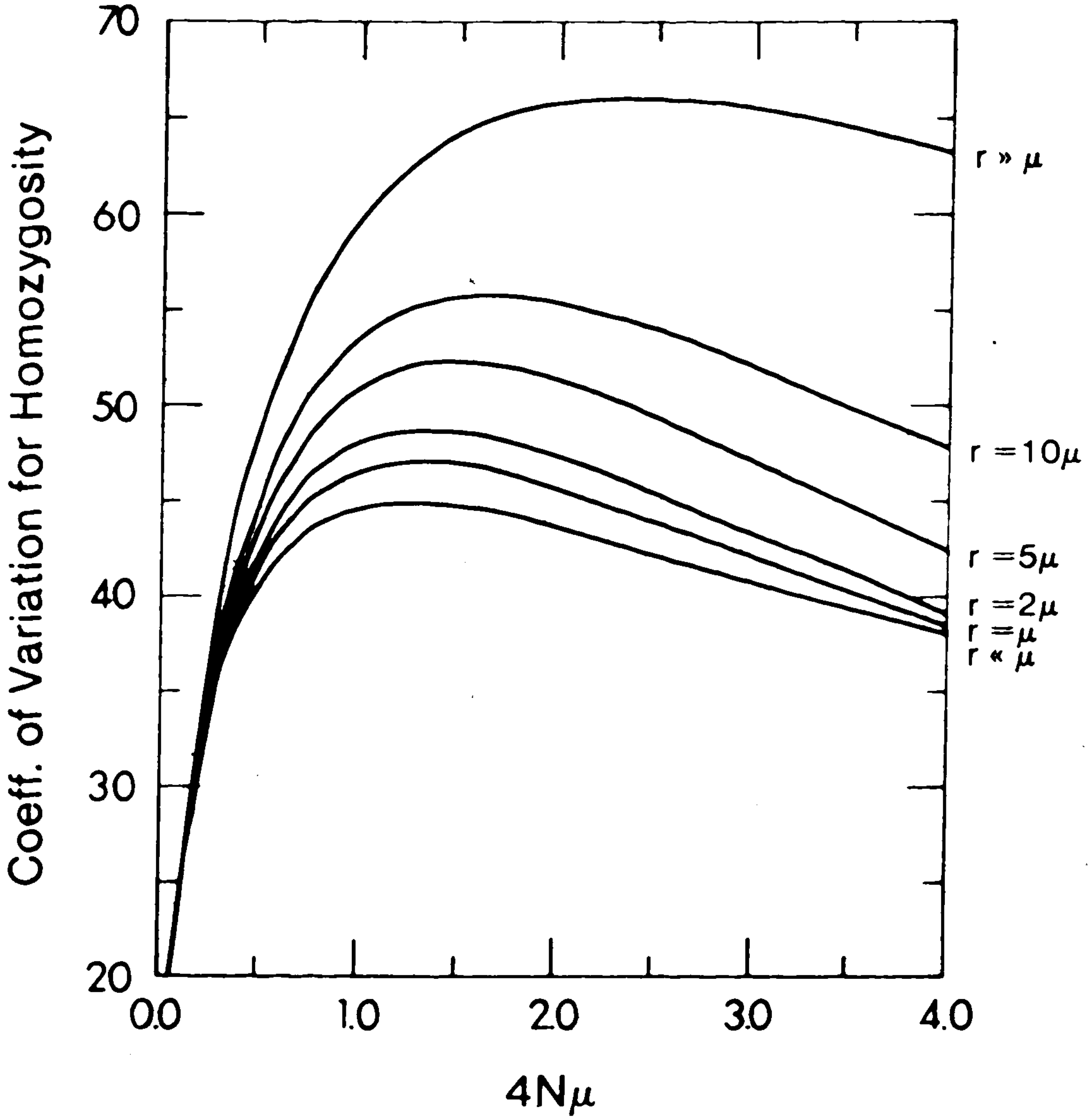
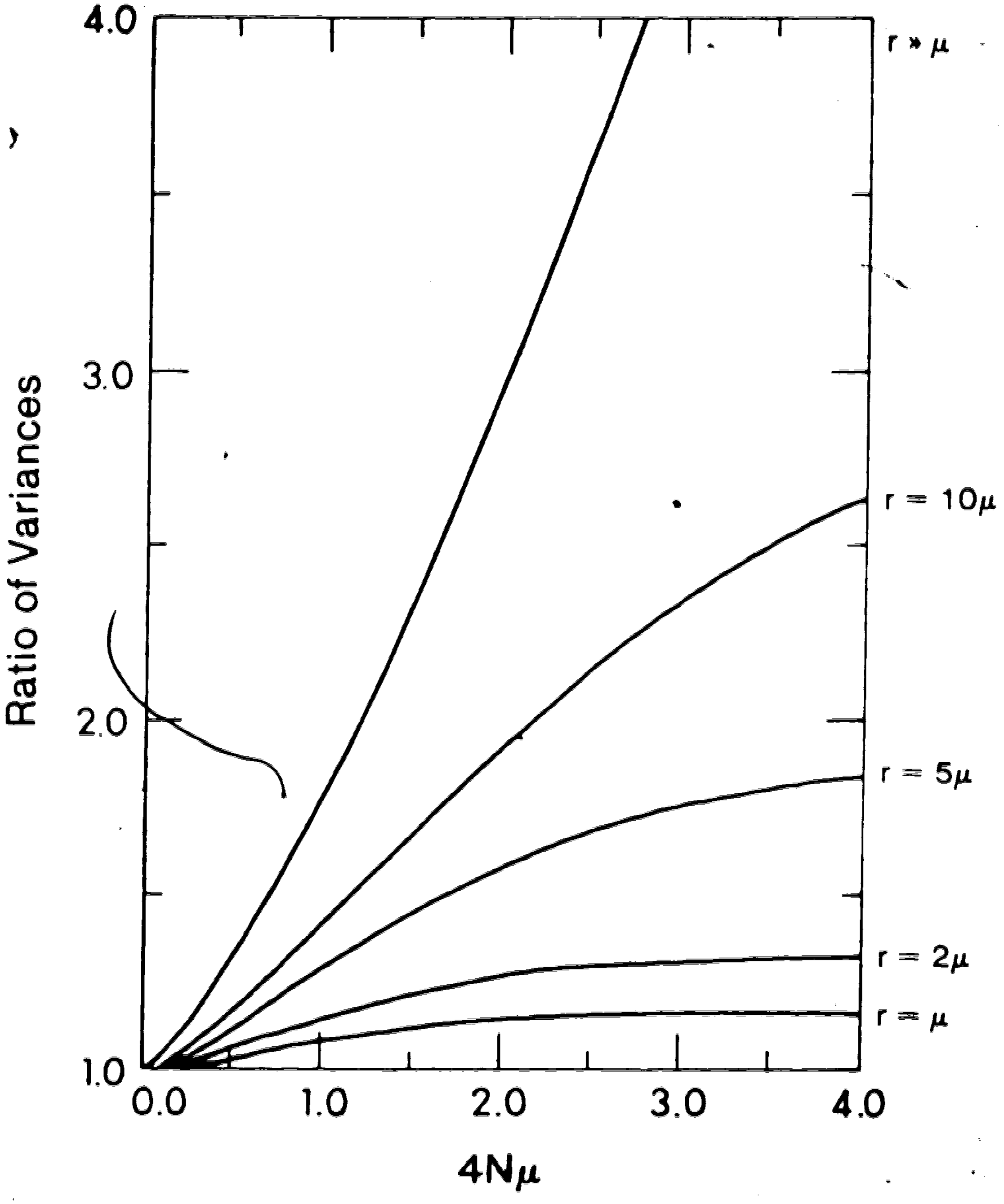

70
60
50
40
30
20
0.0
1.0
2.0
3.0
4.0
Coeff. of Variation for Homozygosity
4Nμ
r » μ
r = 10μ
r = 5μ
r = 2μ
r = μ
r « μ

Figure 6.4: The ratio of the variance of homozygosity of a gene consisting of two sites to the variance of homozygosity of a gene with one site and with the same variability.


4.0
3.0
2.0
1.0
Ratio of Variances
0.0
1.0
2.0
3.0
4.0
4Nμ
r ≫ μ
r = 10μ
r = 5μ
r = 2μ
r = μ

recombination would be to increase the value of $\theta$ to match the increased variability due to recombination. In a sense, trying to find an "effective mutation rate". This can be done by setting $\theta = (1/\phi_{200}) - 1$. This value of $\theta$ gives a single locus the same expected homozygosity. The expected variance of homozygosity for a single locus model with this $\theta$ is

$$\frac{2\theta}{(1+\theta)^2(2+\theta)(3+\theta)}$$

(Watterson, 1974; Stewart, 1976). However, Strobeck and Morgan (1978) argued that the variance of a single locus model even with an increased $\theta$ would underestimate the true variance. In Figure 6.4 the ratio of the true variance of homozygosity to the adjusted variance of a single locus model is compared for $0.0 \leq \theta \leq 4.0$. When the amount of recombination is large, the ratio quickly increases above one as the mutation rate increases. The ratio is small only when $\theta < 0.5$ and r is small. Therefore, intragenic recombination can not be modelled by increasing the mutation rate because the variance and presumably all other moments about the mean are changed.

## Summary

Identity coefficients are used to construct a sufficient set of equations to determine the fourth order moments of gene frequencies for two linked loci. This allows the variance of the expected squared linkage disequilibrium to be found. It is shown that the coefficient of variation is generally greater than one and if the mutation rate is small, the standard deviation is more than four times the size of the mean. This demonstrates that squared linkage disequilibrium is a highly variable quantity. The variance of homozygosity for a gene which consists of two sites can also be obtained. Recombination between these sites increases the variance of homozygosity, suggesting that intragenic recombination significantly changes all the expected moments of gene frequencies if $4N\mu > 1.0$ and $r > \mu$.

# Chapter 7

## Conclusions

Throughout the preceeding chapters the method of identity coefficients has been used to solve several problems. This method is not the only way in which the problems can be solved. Evaluating the moments of the continuous diffusion approximation would give similar answers. Wright's path coefficients will give exactly the same answers. There are probably many different ways in which these problems can be approached. The advantage of the method of identity coefficients (and the other probability methods mentioned in Chapter 1) is that it is simple and intuitive. Indeed, the recursion relationships almost write themselves. When some quantity has been determined to be of interest and it has been defined in terms of a probability, then a recursion relationship can be found for this quantity using simple probability arguments. In writing such a recursion relationship it is often found that other probabilities are required. Recursion relationships for these probabilities can be written and may suggest that still more are required. Eventually a complete set is determined and can then be solved. As stated by Cockerham (1967), "While Malecot's definitions and methods must lead to the same results as does Wright's, they are generally easier to grasp and apply, requiring only simple probability arguments, for those not well versed in path coefficients". The simplicity of this method makes it very useful and the preceeding chapters give only a slight indication of what can be done using identity coefficients.

This method has been used to examine several properties of linkage disequilibrium in Chapters 2, 3 and 6. In general these studies point out a few problems that must be considered. First, higher order linkage disequilibria among several loci need not be a strong indication of the effects of selection. Although smaller than two-locus disequilibria, three-locus disequilibria is of the same order of magnitude. Secondly, the sum of squares of linkage disequilibrium is not a "well-behaved" function in a partially selfing, finite population (Chpt. 3). The linkage disequilibria may increase or decrease with different rates of selfing depending on the mutation rates. This problem can be circumvented by considering the standard squared linkage disequilibrium, a quantity related to the correlation of gene frequencies. However, Chapter 6 shows that there are problems with this quantity as well. The squared standard linkage disequilibrium is a first order Taylor's series approximation to the correlation coefficient. It is shown in Chapter 6 that a second order Taylor's series approximation can be negative when there are rare alleles in the population. This questions the accuracy of the Taylor's series approximation of the squared standard linkage disequilibrium to the correlation coefficient. In this chapter we also demonstrate that a major component of this quantity, the sum of squares of the linkage disequilibrium, has a very large variance (particularly with realistic mutation rates). These chapters have considered only the value of the parameters in the whole population. It is necessary to determine their expected values in a sample but such a theory would be difficult to develop.

The effects of intragenic recombination have been examined in Chapters 2, 4 and 6. These studies demonstrate several properties of this process. Even though intragenic recombination may be rare it can have significant effects in some situations. In hybrid individuals and individuals with an interracial background there is a significant chance that they may have unusual or unique combinations of sites within their genes. It would be preferable to determine the actual number of alleles created in hybrid populations by intragenic recombination but the effective number of alleles is suggestive. The importance of these new alleles depends partly upon their fitnesses. Unfortunately, very little is known about the relative fitnesses of different combinations of sites within genes.

The results in Chapter 6 suggest that intragenic recombination may significantly alter the distribution of gene frequencies. This depends on the sizes of the mutation rates and the amount of recombination between sites. If these two processes are sufficiently large, most of the models in common use will be compromised. For example, Ewens' (1972) method to estimate the parameter $4N\mu$ will give an upwardly biased answer. Since the actual sizes of the recombination and mutation rates are not known precisely, it is not clear how large the effects of intragenic recombination will be. The results in Chapter 2 however, demonstrate that the effects of recombination between more than two sites within a gene may not have to be considered. This study shows that at least, the overall homozygosity is accurately modelled by just a two-site model over a wide range of parameters.

The variance of homozygosity in a structured population is examined in Chapter 5. It is shown here that the variance strongly depends on the amount of migration between subpopulations. The expected variance can however, be accurately estimated using the amount of variability present within the subpopulations. This suggests that the structure of a population, unlike intragenic recombination, alters the distribution of gene frequencies in a simple fashion. This result, along with those of Ewens and Gillespie (1974) and Slatkin (1982), demonstrates that the standard theories for a randomly mating population may be applicable to a strucutured population.

# References

Allard, R.W., G.R. Babbel, M.T. Clegg and A.L. Kahler, 1972 Evidence for coadaptation in *Avena barbata*. Proc. Natl. Acad. Sci. U.S.A. 69:3043-3048.

Allard, R.W., A.L. Kahler and B.S. Weir, 1972 The effect of selection on esterase allozymes in a barley population. Genetics 72:489-503.

Bernstein, F., 1930 Further investigations on the theory of blood groups (German). Zeitschrift fuer induktive Abstammungs- und Verebungslehre 56:233-273.

Brown, A.H.D., 1979 Enzyme polymorphisms in plant populations. Theoret. Pop. Biol. 15:1-42.

Brown, A.H.D., E. Nevo and D. Zohary, 1977 Association of alleles at esterase loci in wild barley *Hordeum spontaneum* L. Nature 268 :430-431.

Cavalli-Sforza, L.L. and W.F. Bodmer, 1971 *The Genetics of Human Populations*. W.H. Freeman and Co., San Francisco.

Chakraborty, R. and M. Nei, 1974 Dynamics of gene differentiation between incompletely isolated populations of unequal sizes. Theoret. Pop. Biol. 5:460-469.

Chevalet, C., M. Gillois and R.F. Nasser, 1977 Identity coefficients in finite populations. I. Evolution of identity coefficients in a random mating diploid dioecious population. Genetics 86:697-713.

Chevalet, C. and M. Gillois, 1977 Estimation of genotypic variance components with dominance in small consanguineous populations. In *Proceedings of the International Conference on Quantitative Genetics* (ed. E. Pollack, O. Kempthorne, T.B. Bailey). Iowa State University Press, Ames, Iowa.

Chevalet, C. and M. Gillois, 1978 Inbreeding depression and heterosis: Expected means and variances among inbred lines and their crosses. Ann. Genet. Sel. Anim. **10**:73-98.

Cotterman, C.W., 1940 A calculus for statistico-genetics. Unpubl. thesis, Ohio State University. Columbus Ohio.

Cockerham, C.C., 1967 Group inbreeding and coancestry. Genetics **56**:89-104.

Cockerham, C.C., 1971 Higher order probability functions of identity of alleles by descent. Genetics **69**:235-246.

Cockerham, C.C. and B.S. Weir, 1968 Sib mating with two linked loci. Genetics **60**:629-640.

Cockerham, C.C. and B.S. Weir, 1973 Descent measures for two loci with some applications. Theoret. Pop. Biol. **4**:300-330.

Crick, F, 1979 Split genes and RNA splicing. Science **204**:264-271.

Crow, J.F. and M. Kimura, 1970 *An Introduction to Population Genetics Theory*. Harper and Row, New York.

Ewens, W.J., 1972 The sampling theory of selectively neutral alleles. Theoret. Pop. Biol. **3**:87-112.

Ewens, W.J., 1974 A note on the sampling theory for infinite alleles and infinite sites models. Theoret. Pop. Biol. **6**:143-148.

Ewens, W.J., 1979 Mathematical Population Genetics. Springer-Verlag, New York.

Ewens, W.J. and J.H. Gillespie, 1974 Some simulation results for the neutral allele model, with interpretations. Theoret. Pop. Biol. 6 :35-57.

Falconer, D.S., 1960 Introduction to Quantitative Genetics. Ronald Press Company, New York.

Felsenstein, J., 1976 The theoretical population genetics of variable selection and migration. Ann. Rev. Genet. 10:253-280.

Freeling, M., 1976 Intragenic recombination in maize: Pollen analysis methods and the effect of parental Adh1+ isoalleles. Genetics 83 :701-717

Gilbert, W., 1978 Why genes in pieces? Nature 271:501.

Gillois, M., 1964 La relation d'identité génétique. Thesis, Faculté des Sciences, Paris.

Haldane, J.B.S., 1939 The equilibrium between mutation and random extinction. Ann. Eugenics 9:400-405.

Haldane, J.B.S., 1950 The association of characters as a result of inbreeding and linkage. Ann. Eugenics 15:15-23.

Haldane, J.B.S. and P. Moshinsky, 1939 Inbreeding in Mendelian populations with special reference to human cousin marriage. Ann. Eugenics 9:321-340.

Hamrick, J.L. and R.W. Allard, 1972 Microgeographical variation in allozyme frequencies in Avena barbata. Proc. Natl. Acad. Sci. U.S.A. 69:2100-2104.

Harris, D.L., 1964 Genotypic covariances between inbred relatives. Genetics 50:1319-1348.

Hill, W.G., 1974a Disequilibrium among several linked neutral genes in finite population. I. Mean changes in disequilibrium. Theoret. Pop. Biol. 5:366-392.

Hill, W.G., 1974b Disequilibrium among several linked neutral genes in finite population. II. Variances and covariances of disequilibrium. Theoret. Pop. Biol. 6:184-198.

Hill, W.G., 1975 Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population. Theoret. Pop. Biol. 8 :117-126.

Hill, W.G., 1976 Non-random association of neutral linked genes in finite populations. In Population Genetics and Ecology (ed. S. Karlin and E. Nevo), pp. 339-376. Academic Press, New York.

Hill, W.G., 1977 Correlation of gene frequencies between neutral linked genes in finite populations. Theoret. Pop. Biol. 11:239-248.

Hill, W.G. and A. Robertson, 1968 Linkage disequilibrium in finite populations. Theoret. Appl. Genet. 38:226-231.

Hunt, W.G. and R.K. Selander, 1973 Biochemical genetics of hybridization in European house mice. Heredity 31:11-33.

Jaquard, A., 1974 The Genetic Structure of Populations. Springer-Verlag, New York.

Kahler, A.L. and R.W. Allard, 1970 Genetics of isozyme variants in barley. I. Esterases. Crop Science 10:444-448.

Kempthorne, O., 1957 *An Introduction to Genetic Statistics*. John Wiley and Sons, Inc., New York.

Kimura, M., 1955 Stochastic processes and distribution of gene frequencies under natural selection. Cold Spring Harbor Symposia on Quantitative Biology 20:33-53.

Kimura, M. and J.F. Crow, 1964 The number of alleles that can be maintained in a finite population. Genetics 49:725-738.

Koch, W., K. Edwards and H. Kossel, 1981 Sequencing of the 16S-23S spacer in a ribosomal RNA operon of *Zea mays* chloroplast DNA reveals two split tRNA genes. Cell 25:203-213.

Koehn, R.K. and W.F. Eanes, 1976 An analysis of allelic diversity in natural populations of *Drosophila*: The correlation of rare alleles with heterozygosity. pp. 377-390. In *Population Genetics and Ecology*. (eds. S. Karlin and E. Nevo). Academic Press, New York.

Langley, C.H., Y.N. Tobari and K. Kojima, 1974 Linkage disequilibrium in natural populations of *Drosophila melanogaster*. Genetics 78 :921-936.

Langley, C.H., K. Ito and R.A. Voelker, 1977 Linkage disequilibrium in natural populations of *Drosophila melanogaster*. Seasonal variation. Genetics 86:447-454.

Lessard, S., 1981 Is the between-population variance negligible in the total variance of heterozygosity? Case of a finite number of loci subject to the infinite-allele model in finite monoecious populations. Theoret. Pop. Biol. 20:394-410.

Lewontin, R.C., 1974 *The Genetic Basis of Evolutionary Change*. Columbia University Press, New York.

Li, C.C., 1978 *First Course in Population Genetics*. Boxwood Press, Pacific Grove, California.

Li, W.H. and M. Nei, 1975 Drift variances of heterozygosity and genetic distance in transient states. Genet. Res. 25:229-248.

Malécot, G., 1948 *Les Mathématiques de l'Hérédité*. Masson et Cie, Paris.

Malécot, G., 1969 *The Mathematics of Heredity*. (trans. by D.M. Yermanos), W.H. Freeman and Co., San Francisco.

Marshall, D.R. and R.W. Allard, 1969 The genetics of electrophoretic variants in *Avena*. I. The esterase $E_4$, $E_9$, $E_{10}$, phosphatase $P_5$ and anodal peroxidase $APX_5$ loci in *A. barbata*. J. Hered. 60:17-19.

Marshall, D.R. and R.W. Allard, 1970 Maintenance of isozyme polymorphisms in natural populations of *Avena barbata*. Genetics 66:393-399.

Maruyama, T., 1970 Effective number of alleles in a subdivided population. Theoret. Pop. Biol. 1:273-306.

Maruyama, T., 1977 *Stochastic Problems in Population Genetics*. Springer-Verlag, New York.

Mayr, E., 1963 *Animal Species and Evolution*. Belknap Press, Cambridge Mass.

McCarron, M., W. Gelbart and A. Chovnick, 1974 Intracistronic mapping of electrophoretic sites in *Drosophila melanogaster*: Fidelity of information transfer by gene conversion. Genetics 76:289-299.

Morgan, K. and C. Strobeck, 1979 Is intragenic recombination a factor in the maintenance of genetic variation in natural populations? Nature 277:383-384.

Mukai, T., T.K. Watanabe and O. Yamaguchi, 1974 The genetic structure of natural populations of Drosophila melanogaster. XII. Linkage disequilibrium in a large local population. Genetics 77:771-793.

Nei, M. and M.W. Feldman, 1972 Identity of genes by descent within and between populations under mutation and migration pressures. Theoret. Pop. Biol. 3:460-465.

Nevo, E., 1978 Genetic variation in natural populations: Patterns and theory. Theoret. Pop. Biol. 13:121-177.

Noble, B. and J.W. Daniel, 1977 Applied Linear Algebra. Prentice-Hall Englewood Cliffs, New Jersey.

Ohno, S., C. Stenius, L. Christian and G. Schipmann, 1969 De novo mutation-like events observed at the 6PGD locus of the Japanese quail, and the principle of polymorphism breeding more polymorphism. Biochem. Genet. 3:417-428.

Ohta, T. and M. Kimura, 1969 Linkage disequilibrium due to random genetic drift. Genet. Res. 13:47-55.

Roughgarden, J., 1979 Theory of Population Genetics and Evolutionary Ecology: An Introduction. MacMillan Publishing Co., Inc., New York.

Sage, D.E. and R.K. Selander, 1979 Hybridization between species of the Rana pipiens complex in central Texas. Evolution 33:1069-1088.

Serant, D., 1974 Linkage and inbreeding coefficients in a finite random mating population. Theoret. Pop. Biol. 6:251-263.

Serant, D., 1976 An application of kinship process to the gene frequencies: Linkage disequilibrium due to random drift in Mendelian genetics with reversible mutation, and in molecular genetics. Theoret. Pop. Biol. 9:1-11.

Serant, D. and M. Villard, 1972 Linerization of crossing-over and mutation in a finite random-mating population. Theoret. Pop. Biol. 3:249-257.

Short, L.L., 1972 Hybridization, taxonomy and avian evolution. Ann. Missouri Bot. Gard. 59:447-453.

Slatkin, M., 1982 Testing neutrality in subdivided populations. Genetics 100:533-545.

Spiess, E.B., 1977 *Genes in Populations*. John Wiley and Sons, Inc., New York.

Stewart, F.M., 1976 Variability in the amount of heterozygosity maintained by neutral mutations. Theoret. Pop. Biol. 9:188-201.

Strobeck, C., 1976 The algebra of recombination. Adv. in Appl. Prob. 8 :27-29.

Strobeck, C. and K. Morgan, 1978 The effect of intragenic recombination on the number of alleles in a finite population. Genetics 88:829-844.

Thompson, J.N. and R.C. Woodruff, 1978 Mutator genes-pacemakers of evolution. Nature 274:317-321.

Tsuno, K., 1981 Studies on mutation at esterase loci in *Drosophila virilis*. I. Spontaneous mutation rates and newly arisen variants. Jpn. J. Genet. 56:155-174.

Vogeli, G., H. Ohkubo, M.E. Sobel, Y. Yamada, I. Pastan and B. de Crombrugghe, 1981 Structure of the promoter for chicken α2 type I collagen gene. Proc. Natl. Acad. Sci. U.S.A. 78:5334-5338.

Wahli, W., I.B. Dawid, T. Wyler, R. Weber and G.U. Ryffel, 1980 Comparative analysis of the structural organization of two closely related vitellogenin genes in *X. laevis*. Cell 20:107-117.

Watt, W.B., 1972 Intragenic recombination as a source of population genetic variability. Am. Nat. 106:737-753.

Watterson, G.A., 1974 Models for the logarithmic species abundance distributions. Theoret. Pop. Biol. 6:217-250.

Weir, B.S., R.W. Allard and A.L. Kahler, 1972 Analysis of complex allozyme polymorphisms in a barley population. Genetics 72:505-523.

Weir, B.S., R.W. Allard and A.L. Kahler, 1974 Further analysis of complex allozyme polymorphisms in a barley population. Genetics 78:911-919.

Weir, B.S. and C. Cockerham, 1974 Behavior of pairs of loci in finite monoecious populations. Theoret. Pop. Biol. 6:323-354.

Wright, S., 1921 Systems of mating. Genetics 6:111-178.

Wright, S., 1922 Coefficients of inbreeding and relationship. Amer. Nat. 56:330-338.

Wright, S., 1951 The genetical structure of populations. Ann. Eugenics 15:323-354.

Wright, S., 1969 *Evolution and the Genetics of Populations*, Vol. II, The Theory of Gene Frequencies. University of Chicago Press, Chicago.

## Appendix 1.

**Recursion Relationships for the Expected Values of the Identity Coefficients Over Replicate Populations for Three Linked Loci.**

$$\phi'_{A/A} = \frac{1}{2N} + (1- \frac{1}{2N} -2\nu_1)\phi_{A/A}$$

$$\phi'_{B/B} = \frac{1}{2N} + (1- \frac{1}{2N} -2\nu_2)\phi_{B/B}$$

$$\phi'_{C/C} = \frac{1}{2N} + (1- \frac{1}{2N} -2\nu_3)\phi_{C/C}$$

$$\phi'_{AB/AB} = \frac{1}{2N} + (1- \frac{1}{2N} -2\nu_1-2\nu_2-2r_{12})\phi_{AB/AB} + 2r_{12}\Gamma_{AB/A/B}$$

$$\phi'_{BC/BC} = \frac{1}{2N} + (1- \frac{1}{2N} -2\nu_2-2\nu_3-2r_{23})\phi_{BC/BC} + 2r_{23}\Gamma_{BC/B/C}$$

$$\phi'_{AC/AC} = \frac{1}{2N} + (1- \frac{1}{2N} -2\nu_1-2\nu_3-2r_{13})\phi_{AC/AC} + 2r_{13}\Gamma_{AC/A/C}$$

$$\Gamma'_{AB/A/B} = \frac{1}{2N}(\phi_{A/A}+\phi_{B/B}+\phi_{AB/AB}) + (1- \frac{2}{2N} -2\nu_1-2\nu_2-r_{12})\Gamma_{AB/A/B} + r_{12}\Delta_{A/A/B/B}$$

$$\Gamma'_{BC/B/C} = \frac{1}{2N}(\phi_{B/B}+\phi_{C/C}+\phi_{BC/BC}) + (1- \frac{3}{2N} -2\nu_2-2\nu_3-r_{23})\Gamma_{BC/B/C} + r_{23}\Delta_{B/B/C/C}$$

$$\Gamma'_{AC/A/C} = \frac{1}{2N}(\phi_{A/A}+\phi_{C/C}+\phi_{AC/AC}) + (1- \frac{3}{2N} -2\nu_1-2\nu_3-r_{13})\Gamma_{AC/A/C} + r_{13}\Delta_{A/A/C/C}$$

$$\Delta'_{A/A/B/B} = \frac{1}{2N}(\phi_{A/A}+\phi_{B/B}+4\Gamma_{AB/A/B}) + (1- \frac{6}{2N} -2\nu_1-2\nu_2)\Delta_{A/A/B/B}$$

$$\Delta'_{B/B/C/C} = \frac{1}{2N}(\phi_{B/B}+\phi_{C/C}+4\Gamma_{BC/B/C}) + (1- \frac{6}{2N} -2\nu_2-2\nu_3)\Delta_{B/B/C/C}$$

$$\Delta'_{A/A/C/C} = \frac{1}{2N}(\phi_{A/A}+\phi_{C/C}+4\Gamma_{AC/A/C}) + (1- \frac{6}{2N} -2\nu_1-2\nu_3)\Delta_{A/A/C/C}$$

$$\phi'_{ABC/ABC} = \frac{1}{2N} + (1-\frac{1}{2N}-2\nu_1-2\nu_2-2\nu_3-2x_1-2x_2-2x_3)\phi_{ABC/ABC} + 2x_1\Gamma_{ABC/AB/C} + 2x_2\Gamma_{ABC/BC/A} + 2x_3\Gamma_{ABC/AC/B}$$

$$\Gamma'_{ABC/AB/C} = \frac{1}{2N}(\phi_{C/C}+\phi_{AB/AB}+\phi_{ABC/ABC}) + (1-\frac{3}{2N}-2\nu_1-2\nu_2-2\nu_3-x_1-x_2-x_3-r_{12})\Gamma_{ABC/AB/C} + x_1\Delta_{AB/AB/C/C}$$

$$+ x_2\Delta_{AB/BC/A/C} + x_3\Delta_{AB/AC/B/C} + r_{12}\Delta_{ABC/A/B/C}$$

$$\Gamma'_{ABC/BC/A} = \frac{1}{2N}(\phi_{A/A}+\phi_{BC/BC}+\phi_{ABC/ABC}) + (1-\frac{3}{2N}-2\nu_1-2\nu_2-2\nu_3-x_1-x_2-x_3-r_{23})\Gamma_{ABC/BC/A} + x_1\Delta_{AB/BC/A/C}$$

$$+ x_2\Delta_{BC/BC/A/A} + x_3\Delta_{BC/AC/A/B} + r_{23}\Delta_{ABC/A/B/C}$$

$$\Gamma'_{ABC/AC/B} = \frac{1}{2N}(\phi_{B/B}+\phi_{AC/AC}+\phi_{ABC/ABC}) + (1-\frac{3}{2N}-2\nu_1-2\nu_2-2\nu_3-x_1-x_2-x_3-r_{13})\Gamma_{ABC/AC/B} + x_1\Delta_{AB/AC/B/C}$$

$$+ x_2\Delta_{BC/AC/A/B} + x_3\Delta_{AC/AC/B/B} + r_{13}\Delta_{ABC/A/B/C}$$

$$\Gamma'_{AB/BC/AC} = \frac{1}{2N}(\phi_{AB/AB}+\phi_{BC/BC}+\phi_{AC/AC}) + (1-\frac{3}{2N}-2\nu_1-2\nu_2-2\nu_3-r_{12}-r_{23}-r_{13})\Gamma_{AB/BC/AC} + r_{12}\Delta_{BC/AC/A/B}$$

$$+ r_{23}\Delta_{AB/AC/B/C} + r_{13}\Delta_{AB/BC/A/C}$$

$$\Delta'_{ABC/A/B/C} = \frac{1}{2N}(\Gamma_{AB/A/B}+\Gamma_{BC/B/C}+\Gamma_{AC/A/C}+\Gamma_{ABC/AB/C}+\Gamma_{ABC/BC/A}+\Gamma_{ABC/AC/B}) + (1-\frac{6}{2N}-2\nu_1-2\nu_2-2\nu_3-x_1-x_2-x_3)\Delta_{ABC/A/B/C}$$

$$+ x_1\Lambda_{AB/A/B/C/C} + x_2\Lambda_{BC/A/A/B/C} + x_3\Lambda_{AC/A/B/B/C}$$

$$\Delta'_{AB/AB/C/C} = \frac{1}{2N}(\phi_{C/C}+\phi_{AB/AB}+4\Gamma_{ABC/AB/C}) + (1-\frac{6}{2N}-2\nu_1-2\nu_2-2\nu_3-2r_{12})\Delta_{AB/AB/C/C} + 2r_{12}\Lambda_{AB/A/B/C/C}$$

$$\Delta'_{AB/BC/A/C} = \frac{1}{2N}(\Gamma_{AB/A/B}+\Gamma_{BC/B/C}+\Gamma_{AC/A/C}+\Gamma_{ABC/AB/C}+\Gamma_{ABC/BC/A}+\Gamma_{AB/BC/AC}) + (1-\frac{6}{2N}-2\nu_1-2\nu_2-2\nu_3-r_{12}-r_{23})\Delta_{AB/BC/A/C}$$

$$+ r_{12}\Lambda_{BC/A/A/B/C} + r_{23}\Lambda_{AB/A/B/C/C}$$

$$\Delta'_{AB/AC/B/C} = \frac{1}{2N}(\Gamma_{AB/A/B}+\Gamma_{BC/B/C}+\Gamma_{AC/A/C}+\Gamma_{ABC/AB/C}+\Gamma_{ABC/AC/B}+\Gamma_{AB/BC/AC}) + (1-\frac{6}{2N}-2\nu_1-2\nu_2-2\nu_3-r_{12}-r_{13})\Delta_{AB/AC/B/C}$$

$$+ r_{12}\Lambda_{AC/A/B/B/C} + r_{13}\Lambda_{AB/A/B/C/C}$$

$$\Delta'_{BC/BC/A/A} = \frac{1}{2N}(\Phi_{A/A}+\Phi_{BC/BC}+4\Gamma_{ABC/BC/A}) + (1-\frac{6}{2N}-2\nu_1-2\nu_2-2\nu_3-2r_{23})\Delta_{BC/BC/A/A} + 2r_{23}\Lambda_{BC/A/A/B/C}$$

$$\Delta'_{BC/AC/A/B} = \frac{1}{2N}(\Gamma_{AB/A/B}+\Gamma_{BC/B/C}+\Gamma_{AC/A/C}+\Gamma_{ABC/BC/A}+\Gamma_{ABC/AC/B}+\Gamma_{AB/BC/AC}) + (1-\frac{6}{2N}-2\nu_1-2\nu_2-2\nu_3-r_{23}-r_{13})\Delta_{BC/AC/A/B}$$

$$+ r_{23}\Lambda_{AC/A/B/B/C} + r_{13}\Lambda_{BC/A/A/B/C}$$

$$\Delta'_{AC/AC/B/B} = \frac{1}{2N}(\Phi_{B/B}+\Phi_{AC/AC}+4\Gamma_{ABC/AC/B}) + (1-\frac{6}{2N}-2\nu_1-2\nu_2-2\nu_3-2r_{13})\Delta_{AC/AC/B/B} + 2r_{13}\Lambda_{AC/A/B/B/C}$$

$$\Lambda'_{AB/A/B/C/C} = \frac{1}{2N}(\Gamma_{AB/A/B}+\Delta_{B/B/C/C}+\Delta_{A/A/C/C}+2\Delta_{ABC/A/B/C}+\Delta_{AB/AB/C/C}+2\Delta_{AB/BC/A/C}+2\Delta_{AB/AC/B/C})$$

$$+ (1-\frac{10}{2N}-2\nu_1-2\nu_2-2\nu_3-r_{12})\Lambda_{AB/A/B/C/C} + r_{12}\Psi_{A/A/B/B/C/C}$$

$$\Lambda'_{BC/A/A/B/C} = \frac{1}{2N}(\Gamma_{BC/B/C}+\Delta_{A/A/B/B}+\Delta_{A/A/C/C}+2\Delta_{ABC/A/B/C}+2\Delta_{AB/BC/A/C}+\Delta_{BC/BC/A/A}+2\Delta_{BC/AC/A/B})$$

$$+ (1-\frac{10}{2N}-2\nu_1-2\nu_2-2\nu_3-r_{23})\Lambda_{BC/A/A/B/C} + r_{23}\Psi_{A/A/B/B/C/C}$$

$$\Lambda'_{AC/A/B/B/C} = \frac{1}{2N}(\Gamma_{AC/A/C}+\Delta_{A/A/B/B}+\Delta_{B/B/C/C}+2\Delta_{ABC/A/B/C}+2\Delta_{AB/AC/B/C}+2\Delta_{BC/AC/A/B}+\Delta_{AC/AC/B/B})$$

$$+ (1-\frac{10}{2N}-2\nu_1-2\nu_2-2\nu_3-r_{13})\Lambda_{AC/A/B/B/C} + r_{13}\Psi_{A/A/B/B/C/C}$$

$$\Psi'_{A/A/B/B/C/C} = \frac{1}{2N}(\Delta_{A/A/B/B}+\Delta_{B/B/C/C}+\Delta_{A/A/C/C}+4\Lambda_{AB/A/B/C/C}+4\Lambda_{BC/A/A/B/C}+4\Lambda_{AC/A/B/B/C}) + (1-\frac{15}{2N}-2\nu_1-2\nu_2-2\nu_3)\Psi_{A/A/B/B/C/C}$$

# Appendix 2.

## Recursion Relationships for the Expected Values of the Identity Coefficients Over Replicate Populations for a Partially Selfing, Finite Population.

$$\Psi_{(A/A)}' = (1-\mu)^2\Big(S(\tfrac{1}{2}+\tfrac{1}{2}\Psi_{(A/A)}) + (1-S)\Phi_{(A)(A)}\Big)$$

$$\Psi_{(B/B)}' = (1-\nu)^2\Big(S(\tfrac{1}{2}+\tfrac{1}{2}\Psi_{(B/B)}) + (1-S)\Phi_{(B)(B)}\Big)$$

$$\Phi_{(A)(A)}' = (1-\mu)^2\Big(\frac{1}{N}(\tfrac{1}{2}+\tfrac{1}{2}\Psi_{(A/A)}) + (1-\frac{1}{N})\Phi_{(A)(A)}\Big)$$

$$\Phi_{(B)(B)}' = (1-\nu)^2\Big(\frac{1}{N}(\tfrac{1}{2}+\tfrac{1}{2}\Psi_{(B/B)}) + (1-\frac{1}{N})\Phi_{(B)(B)}\Big)$$

$$\Psi_{(AB/AB)}' = (1-\mu)^2(1-\nu)^2\Big(S[(1-r)^2\Lambda_1 + 2r(1-r)\Lambda_2 + r^2\Lambda_1] + (1-S)\Phi_{(AB)(AB)}\Big)$$

$$\Phi_{(AB)(AB)}' = (1-\mu)^2(1-\nu)^2\Big((1-r)^2[\frac{1}{N}\Lambda_1 + (1-\frac{1}{N})\Phi_{(AB)(AB)}] + 2r(1-r)[\frac{1}{N}\Lambda_2 + (1-\frac{1}{N})\Phi_{(AB)(A/B)}]$$

$$+ r^2[\frac{1}{N}\Lambda_1 + (1-\frac{1}{N})\Phi_{(A/B)(A/B)}]\Big)$$

$$\Phi_{(AB)(A/B)}' = (1-\mu)^2(1-\nu)^2\Big(S\{(1-r)[\frac{1}{\text{[illegible]}} + (1-\frac{1}{N})\Omega_1] + r[\frac{1}{N}\Lambda_3 + (1-\frac{1}{N})\Omega_2]\}$$

$$+ (1-S)\{(1-r)[\frac{1}{N}\Omega_3 + \text{[illegible]} + (1-\frac{2}{N})\Gamma_{(AB)(A)(B)}]$$

$$+ r[\frac{1}{N}\Omega_3 + \frac{1}{N}\Omega_4 + (1-\frac{2}{N})\Gamma_{(A/B)(A)(B)}]\}\Big)$$

$$\Phi_{(AB/B)(A)}' = (1-\mu)^2(1-\nu)^2\Big(S[\frac{1}{N}\Lambda_3 + (1-\frac{1}{N})\Omega_4] + (1-S)\{(1-r)[\frac{1}{N}\Omega_1 + \frac{1}{N}\Omega_3 + (1-\frac{2}{N})\Gamma_{(AB)(A)(B)}]$$

$$+ r[\frac{1}{N}\Omega_2 + \frac{1}{N}\Omega_3 + (1-\frac{2}{N})\Gamma_{(A/B)(A)(B)}]\}\Big)$$

$$\Phi_{(AB/A)(B)}' = (1-\mu)^2(1-\nu)^2\Big(S[\frac{1}{N}\Lambda_3 + (1-\frac{1}{N})\Omega_3] + (1-S)\{(1-r)[\frac{1}{N}\Omega_1 + \frac{1}{N}\Omega_4 + (1-\frac{2}{N})\Gamma_{(AB)(A)(B)}]$$

$$+ r[\frac{1}{N}\Omega_2 + \frac{1}{N}\Omega_4 + (1-\frac{2}{N})\Gamma_{(A/B)(A)(B)}]\}\Big)$$

$$\Phi_{(A/A)(B/B)}' = (1-\mu)^2(1-\nu)^2\Big\{S^2[\tfrac{1}{N}\Lambda_3 + (1-\tfrac{1}{N})\Omega_5] + S(1-S)[\tfrac{2}{N}\Phi_{(AB/A)(B)} + (1-\tfrac{2}{N})\Gamma_{(A/A)(B)(B)}]$$

$$+ S(1-S)[\tfrac{2}{N}\Phi_{(AB/B)(A)} + (1-\tfrac{2}{N})\Gamma_{(B/B)(A)(A)}]$$

$$+ (1-S)^2[\frac{2}{N(N-1)}\Omega_6 + \frac{4(N-2)}{N(N-1)}\Pi_3 + \frac{(N-2)(N-3)}{N(N-1)}\Delta_{(A)(B)(A)(B)}]\Big\}$$

$$\Phi_{(A/B)(A/B)}' = (1-\mu)^2(1-\nu)^2\Big\{S^2[\tfrac{1}{N}\Lambda_3 + (1-\tfrac{1}{N})\Omega_6] + 2S(1-S)[\tfrac{1}{N}\Omega_3 + \tfrac{1}{N}\Omega_4 + (1-\tfrac{2}{N})\Gamma_{(A/B)(A)(B)}]$$

$$+ (1-S)^2[\frac{1}{N(N-1)}(\Omega_5 + \Omega_6) + \frac{N-2}{N(N-1)}(\Pi_1 + \Pi_2 + 2\Pi_3) + \frac{(N-2)(N-3)}{N(N-1)}\Delta_{(A)(B)(A)(B)}]\Big\}$$

$$\Gamma_{(AB)(A)(B)}' = (1-\mu)^2(1-\nu)^2\Big\{(1-r)[\frac{1}{N^2}\Lambda_3 + \frac{N-1}{N^2}(\Omega_1 + \Omega_3 + \Omega_4) + \frac{(N-1)(N-2)}{N^2}\Gamma_{(AB)(A)(B)}]$$

$$+ r[\frac{1}{N^2}\Lambda_3 + \frac{N-1}{N^2}(\Omega_2 + \Omega_3 + \Omega_4) + \frac{(N-1)(N-2)}{N^2}\Gamma_{(A/B)(A)(B)}]\Big\}$$

$$\Gamma_{(B/B)(A)(A)}' = (1-\mu)^2(1-\nu)^2\Big\{S[\frac{1}{N^2}\Lambda_3 + \frac{N-1}{N}(2\Omega_4 + \Omega_5) + \frac{(N-1)(N-2)}{N^2}\Pi_2]$$

$$+ (1-S)[\frac{1}{N^2}(\Omega_1 + \Omega_2 + 2\Omega_3) + \frac{N-2}{N^2}(\Pi_1 + 4\Pi_3) + \frac{(N-2)(N-3)}{N^2}\Delta_{(A)(B)(A)(B)}]\Big\}$$

$$\Gamma_{(A/A)(B)(B)}' = (1-\mu)^2(1-\nu)^2\Big\{S[\frac{1}{N^2}\Lambda_3 + \frac{N-1}{N^2}(2\Omega_3 + \Omega_5) + \frac{(N-1)(N-2)}{N^2}\Pi_1]$$

$$+ (1-S)[\frac{1}{N^2}(\Omega_1 + \Omega_2 + 2\Omega_4) + \frac{N-2}{N^2}(\Pi_2 + 4\Pi_3) + \frac{(N-2)(N-3)}{N^2}\Delta_{(A)(B)(A)(B)}]\Big\}$$

$$\Gamma_{(A/B)(A)(B)}' = (1-\mu)^2(1-\nu)^2\Big\{S[\frac{1}{N^2}\Lambda_3 + \frac{N-1}{N^2}(\Omega_3 + \Omega_4 + \Omega_5) + \frac{(N-1)(N-2)}{N^2}\Pi_3]$$

$$+ (1-S)[\frac{1}{N^2}(\Omega_3 + \Omega_4 + \Omega_5 + \Omega_6) + \frac{N-2}{N^2}(\Pi_1 + \Pi_2 + 3\Pi_3) + \frac{(N-2)(N-3)}{N^2}\Delta_{(A)(B)(A)(B)}]\Big\}$$

$$\Delta_{(A)(B)(A)(B)}' = (1-\mu)^2(1-\nu)^2\Big\{\frac{1}{N^2}\Lambda_3 + \frac{N-1}{N^2}(2\Omega_3 + 2\Omega_4 + \Omega_5 + 2\Omega_6) + \frac{(N-1)(N-2)}{N^2}(\Pi_1 + \Pi_2 + 4\Pi_3)$$

$$+ \frac{(N-1)(N-2)(N-3)}{N^2}\Delta_{(A)(B)(A)(B)}\Big\}$$

where $\Lambda_1 = \tfrac{1}{2} + \tfrac{1}{2}\Psi_{(AB/AB)}$

$\Lambda_2 = \tfrac{1}{2}\Psi_{(A/A)} + \tfrac{1}{2}\Psi_{(B/B)}$

$\Lambda_3 = \tfrac{1}{4} + \tfrac{1}{4}\Psi_{(A/A)} + \tfrac{1}{4}\Psi_{(B/B)} + \tfrac{1}{4}\Psi_{(AB/AB)}$

$$\Omega_1 = \tfrac{1}{2}\Phi_{(AB)(AB)} + \tfrac{1}{2}\Phi_{(AB)(A/B)}$$

$$\Omega_2 = \tfrac{1}{2}\Phi_{(AB)(A/B)} + \tfrac{1}{2}\Phi_{(A/B)(A/B)}$$

$$\Omega_3 = \tfrac{1}{2}\Phi_{(B)(B)} + \tfrac{1}{2}\Phi_{(AB/A)(B)}$$

$$\Omega_4 = \tfrac{1}{2}\Phi_{(A)(A)} + \tfrac{1}{2}\Phi_{(AB/B)(A)}$$

$$\Omega_5 = \tfrac{1}{4} + \tfrac{1}{4}\Psi_{(A/A)} \overset{\Psi}{+} \tfrac{1}{4}\Psi_{(B/B)} + \tfrac{1}{4}\Phi_{(A/A)(B/B)}$$

$$\Omega_6 = \tfrac{1}{4}\Phi_{(AB)(AB)} + \tfrac{1}{2}\Phi_{(AB)(A/B)} + \tfrac{1}{4}\Phi_{(A/B)(A/B)}$$

$$R_1 = \tfrac{1}{2}\Phi_{(B)(B)} + \tfrac{1}{2}\Gamma_{(A/A)(B)(B)}$$

$$R_2 = \tfrac{1}{2}\Phi_{(A)(A)} + \tfrac{1}{2}\Gamma_{(B/B)(A)(A)}$$

$$R_3 = \tfrac{1}{2}\Gamma_{(AB)(A)(B)} + \tfrac{1}{2}\Gamma_{(A/B)(A)(B)}$$

If $N>>1$ , $\mu=O(\frac{1}{N})$ , $\nu=O(\frac{1}{N})$ , and $r=O(\frac{1}{N})$ , then these equations can be approximated by

$$\begin{aligned}
\Psi_{(A/A)}{}' &= S(\tfrac{1}{2}+\tfrac{1}{2}\Psi_{(A/A)}) + (1-S)\Phi_{(A)(A)} \\
\Psi_{(B/B)}{}' &= S(\tfrac{1}{2}+\tfrac{1}{2}\Psi_{(B/B)}) + (1-S)\Phi_{(B)(B)} \\
\Psi_{(AB/AB)}{}' &= S(\tfrac{1}{2}+\tfrac{1}{2}\Psi_{(AB/AB)}) + (1-S)\Phi_{(AB)(AB)} \\
\Phi_{(AB)(A/B)}{}' &= S(\tfrac{1}{2}\Phi_{(AB)(AB)}+\tfrac{1}{2}\Phi_{(AB)(A/B)}) + (1-S)\Gamma_{(AB)(A)(B)} \\
\Phi_{(AB/B)(A)}{}' &= S(\tfrac{1}{2}\Phi_{(A)(A)}+\tfrac{1}{2}\Phi_{(AB/B)(A)}) + (1-S)\Gamma_{(AB)(A)(B)} \\
\Phi_{(AB/A)(B)}{}' &= S(\tfrac{1}{2}\Phi_{(B)(B)}+\tfrac{1}{2}\Phi_{(AB/A)(B)}) + (1-S)\Gamma_{(AB)(A)(B)} \\
\Phi_{(A/A)(B/B)}{}' &= S^2(\tfrac{1}{4}+\tfrac{1}{4}\Psi_{(A/A)}+\tfrac{1}{4}\Psi_{(B/B)}+\tfrac{1}{4}\Phi_{(A/A)(B/B)}) + S(1-S)(\Gamma_{(A/A)(B)(B)}+\Gamma_{(B/B)(A)(A)}) \\
&\quad + (1-S)^2\Delta_{(A)(B)(A)(B)} \\
\Phi_{(A/B)(A/B)}{}' &= S^2(\tfrac{1}{4}\Phi_{(AB)(AB)}+\tfrac{1}{2}\Phi_{(AB)(A/B)}+\tfrac{1}{4}\Phi_{(A/B)(A/B)}) + 2S(1-S)\Gamma_{(A/B)(A)(B)} \\
&\quad + (1-S)^2\Delta_{(A)(B)(A)(B)}
\end{aligned} \qquad \text{(A1)}$$

$$\Gamma_{(B/B)(A)(A)}{}' = S(\tfrac{1}{2}\phi_{(A)(A)}+\tfrac{1}{2}\Gamma_{(B/B)(A)(A)}) + (1-S)\Delta_{(A)(B)(A)(B)}$$

$$\Gamma_{(A/A)(B)(B)}{}' = S(\tfrac{1}{2}\phi_{(B)(B)}+\tfrac{1}{2}\Gamma_{(A/A)(B)(B)}) + (1-S)\Delta_{(A)(B)(A)(B)}$$

$$\Gamma_{(A/B)(A)(B)}{}' = S(\tfrac{1}{2}\Gamma_{(AB)(A)(B)}+\tfrac{1}{2}\Gamma_{(A/B)(A)(B)}) + (1-S)\Delta_{(A)(B)(A)(B)}$$

neglecting terms of $O(\frac{1}{N})$ or less and

$$\phi_{(A)(A)}{}' = \frac{1}{N}(\tfrac{1}{2}+\tfrac{1}{2}\psi_{(A/A)}) + (1-\frac{1}{N}-2\mu)\phi_{(A)(A)}$$

$$\phi_{(B)(B)}{}' = \frac{1}{N}(\tfrac{1}{2}+\tfrac{1}{2}\psi_{(B/B)}) + (1-\frac{1}{N}-2\nu)\phi_{(B)(B)}$$

$$\phi_{(AB)(AB)}{}' = \frac{1}{N}(\tfrac{1}{2}+\tfrac{1}{2}\psi_{(AB/AB)}) + (1-\frac{1}{N}-2\mu-2\nu-2r)\phi_{(AB)(AB)} + 2r\phi_{(AB)(A/B)}$$

$$\tag{A2}$$

$$\Gamma_{(AB)(A)(B)}{}' = \frac{1}{N}(\tfrac{1}{2}\phi_{(A)(A)}+\tfrac{1}{2}\phi_{(AB/B)(A)}+\tfrac{1}{2}\phi_{(B)(B)}+\tfrac{1}{2}\phi_{(AB/A)(B)}+\tfrac{1}{2}\phi_{(AB)(AB)}+\tfrac{1}{2}\phi_{(AB)(A/B)})$$

$$+ (1-\frac{3}{N}-2\mu-2\nu-r)\Gamma_{(AB)(A)(B)} + r\Gamma_{(A/B)(A)(B)}$$

$$\Delta_{(A)(B)(A)(B)}{}' = \frac{1}{N}(\tfrac{1}{2}\phi_{(A)(A)}+\tfrac{1}{2}\Gamma_{(B/B)(A)(A)}+\tfrac{1}{2}\phi_{(B)(B)}+\tfrac{1}{2}\Gamma_{(A/A)(B)(B)}+2\Gamma_{(AB)(A)(B)}+2\Gamma_{(A/B)(A)(B)})$$

$$+ (1-\frac{6}{N}-2\mu-2\nu)\Delta_{(A)(B)(A)(B)}$$

neglecting terms of $O(\frac{1}{N^2})$ or less.

# Appendix 3.

## Recursion Relationships for the Expected Values of the Identity Coefficients Over Replicate Populations for Two Loci with Two Subpopulations.

$$\gamma'_{(a)_1(a)_1} = \frac{1}{2N_1} + (1 - \frac{1}{2N_1} - 2m_1 - 2v_1)\gamma_{(a)_1(a)_1} + 2m_1\gamma_{(a)_1(a)_2}$$

$$\gamma'_{(a)_1(a)_2} = (1 - m_1 - m_2 - 2v_1)\gamma_{(a)_1(a)_2} + m_1\gamma_{(a)_2(a)_2} + m_2\gamma_{(a)_1(a)_1}$$

$$\gamma'_{(a)_2(a)_2} = \frac{1}{2N_2} + (1 - \frac{1}{2N_2} - 2m_2 - 2v_1)\gamma_{(a)_2(a)_2} + 2m_2\gamma_{(a)_1(a)_2}$$

$$\gamma'_{(b)_1(b)_1} = \frac{1}{2N_1} + (1 - \frac{1}{2N_1} - 2m_1 - 2v_2)\gamma_{(b)_1(b)_1} + 2m_1\gamma_{(b)_1(b)_2}$$

$$\gamma'_{(b)_1(b)_2} = (1 - m_1 - m_2 - 2v_2)\gamma_{(b)_1(b)_2} + m_1\gamma_{(b)_2(b)_2} + m_2\gamma_{(b)_1(b)_1}$$

$$\gamma'_{(b)_2(b)_2} = \frac{1}{2N_2} + (1 - \frac{1}{2N_2} - 2m_2 - 2v_2)\gamma_{(b)_2(b)_2} + 2m_2\gamma_{(b)_1(b)_2}$$

$$\phi'_{(ab)_1(ab)_1} = \frac{1}{2N_1} + (1 - \frac{1}{2N_1} - 2r - 2m_1 - 2v_1 - 2v_2)\phi_{(ab)_1(ab)_1} + 2r\Gamma_{(ab)_1(a)_1(b)_1} + 2m_1\phi_{(ab)_1(ab)_2}$$

$$\phi'_{(ab)_1(ab)_2} = (1 - 2r - m_1 - m_2 - 2v_1 - 2v_2)\phi_{(ab)_1(ab)_2} + r\Gamma_{(ab)_2(a)_1(b)_1} + r\Gamma_{(ab)_1(a)_2(b)_2} + m_1\phi_{(ab)_2(ab)_2} + m_2\phi_{(ab)_1(ab)_1}$$

$$\phi'_{(ab)_2(ab)_2} = \frac{1}{2N_2} + (1 - \frac{1}{2N_2} - 2r - 2m_2 - 2v_1 - 2v_2)\phi_{(ab)_2(ab)_2} + 2r\Gamma_{(ab)_2(a)_2(b)_2} + 2m_2\phi_{(ab)_1(ab)_2}$$

$$\dot{\Gamma}_{(ab)_1(a)_1(b)_1} = \frac{1}{2N_1}(\Phi_{(a)_1(a)_1}+\Phi_{(b)_1(b)_1}+\Phi_{(ab)_1(ab)_1}) + (1-\frac{3}{2N_1}-r-3m_1-2v_1-2v_2)\Gamma_{(ab)_1(a)_1(b)_1} + r\Delta_{(a)_1(b)_1(a)_1(b)_1} + m_1\Gamma_{(ab)_2(a)_1(b)_1}$$
$$+ m_1\Gamma_{(ab)_1(a)_1(b)_2} + m_1\Gamma_{(ab)_1(a)_2(b)_1}$$

$$\dot{\Gamma}_{(ab)_1(a)_1(b)_2} = \frac{1}{2N_1}\Phi_{(b)_1(b)_2} + (1-\frac{1}{2N_1}-r-2m_1-m_2-2v_1-2v_2)\Gamma_{(ab)_1(a)_1(b)_2} + r\Delta_{(a)_1(b)_1(a)_1(b)_2} + m_1\Gamma_{(ab)_1(a)_2(b)_2} + m_1\Gamma_{(ab)_2(a)_1(b)_2}$$
$$+ m_2\Gamma_{(ab)_1(a)_1(b)_1}$$

$$\dot{\Gamma}_{(ab)_1(a)_2(b)_1} = \frac{1}{2N_1}\Phi_{(a)_1(a)_2} + (1-\frac{1}{2N_1}-r-2m_1-m_2-2v_1-2v_2)\Gamma_{(ab)_1(a)_2(b)_1} + r\Delta_{(a)_1(b)_1(a)_2(b)_1} + m_1\Gamma_{(ab)_1(a)_2(b)_2} + m_1\Gamma_{(ab)_2(a)_2(b)_1}$$
$$+ m_2\Gamma_{(ab)_1(a)_1(b)_1}$$

$$\dot{\Gamma}_{(ab)_2(a)_1(b)_1} = \frac{1}{2N_1}\Phi_{(ab)_1(ab)_2} + (1-\frac{1}{2N_1}-r-2m_1-m_2-2v_1-2v_2)\Gamma_{(ab)_2(a)_1(b)_1} + r\Delta_{(a)_1(b)_1(a)_2(b)_2} + m_1\Gamma_{(ab)_2(a)_1(b)_2} + m_1\Gamma_{(ab)_2(a)_2(b)_1}$$
$$+ m_2\Gamma_{(ab)_1(a)_1(b)_1}$$

$$\dot{\Gamma}_{(ab)_1(a)_2(b)_2} = \frac{1}{2N_2}\Phi_{(ab)_1(ab)_2} + (1-\frac{1}{2N_2}-r-m_1-2m_2-2v_1-2v_2)\Gamma_{(ab)_1(a)_2(b)_2} + r\Delta_{(a)_1(b)_1(a)_2(b)_2} + m_1\Gamma_{(ab)_2(a)_2(b)_2} + m_2\Gamma_{(ab)_1(a)_1(b)_2},$$
$$+ m_2\Gamma_{(ab)_1(a)_2(b)_1}$$

$$\dot{\Gamma}_{(ab)_2(a)_1(b)_2} = \frac{1}{2N_2}\Phi_{(a)_1(a)_2} + (1-\frac{1}{2N_2}-r-m_1-2m_2-2v_1-2v_2)\Gamma_{(ab)_2(a)_1(b)_2} + r\Delta_{(a)_1(b)_2(a)_2(b)_2} + m_1\Gamma_{(ab)_2(a)_2(b)_2} + m_2\Gamma_{(ab)_1(a)_1(b)_2}$$
$$+ m_2\Gamma_{(ab)_2(a)_1(b)_1}$$

$$\dot{\Gamma}_{(ab)_2(a)_2(b)_1} = \frac{1}{2N_2}\gamma_{(b)_1(b)_2} + (1 - \frac{1}{2N_2} - r - m_1 - 2m_2 - 2v_1 - 2v_2)\Gamma_{(ab)_2(a)_2(b)_1} + r\Delta_{(a)_2(b)_1(a)_2(b)_2} + m_1\Gamma_{(ab)_2(a)_2(b)_2} + m_2\Gamma_{(ab)_1(a)_2(b)_1}$$

$$+ m_2\Gamma_{(ab)_2(a)_1(b)_1}$$

$$\dot{\Gamma}_{(ab)_2(a)_2(b)_2} = \frac{1}{2N_2}(\gamma_{(a)_2(a)_2} + \gamma_{(b)_2(b)_2} + \gamma_{(ab)_2(ab)_2}) + (1 - \frac{2}{2N_2} - r - 2m_2 - 2v_1 - 2v_2)\Gamma_{(ab)_2(a)_2(b)_2} + r\Delta_{(a)_2(b)_2(a)_2(b)_2} + m_2\Gamma_{(ab)_1(a)_2(b)_2}$$

$$+ m_2\Gamma_{(ab)_2(a)_1(b)_2} + m_2\Gamma_{(ab)_2(a)_2(b)_1}$$

$$\dot{\Delta}_{(a)_1(b)_1(a)_1(b)_1} = \frac{1}{2N_1}(\gamma_{(a)_1(a)_1} + \gamma_{(b)_1(b)_1} + 4\Gamma_{(ab)_1(a)_1(b)_1}) + (1 - \frac{6}{2N_1} - 4m_1 - 2v_1 - 2v_2)\Delta_{(a)_1(b)_1(a)_1(b)_1} + 2m_1\Delta_{(a)_1(b)_1(a)_2(b)_1} + 2m_1\Delta_{(a)_1(b)_1(a)_1(b)_2}$$

$$\dot{\Delta}_{(a)_1(b)_1(a)_1(b)_2} = \frac{1}{2N_1}(\gamma_{(a)_1(a)_2} + 2\Gamma_{(ab)_1(a)_1(b)_2}) + (1 - \frac{3}{2N_1} - 3m_1 - m_2 - 2v_1 - 2v_2)\Delta_{(a)_1(b)_1(a)_1(b)_2} + 2m_1\Delta_{(a)_1(b)_1(a)_2(b)_2} + m_1\Delta_{(a)_1(b)_2(a)_1(b)_2}$$

$$+ m_2\Delta_{(a)_1(b)_1(a)_1(b)_1}$$

$$\dot{\Delta}_{(a)_1(b)_1(a)_2(b)_1} = \frac{1}{2N_1}(\gamma_{(a)_1(a)_2} + 2\Gamma_{(ab)_1(a)_2(b)_1}) + (1 - \frac{3}{2N_1} - 3m_1 - m_2 - 2v_1 - 2v_2)\Delta_{(a)_1(b)_1(a)_2(b)_1} + 2m_1\Delta_{(a)_1(b)_1(a)_2(b)_2} + m_1\Delta_{(a)_2(b)_1(a)_2(b)_1}$$

$$+ m_2\Delta_{(a)_1(b)_1(a)_1(b)_1}$$

$$\dot{\Delta}_{(a)_1(b)_2(a)_1(b)_2} = \frac{1}{2N_1}\gamma_{(b)_2(b)_2} + \frac{1}{2N_2}\gamma_{(a)_1(a)_1} + (1 - \frac{1}{2N_1} - \frac{1}{2N_2} - 2m_1 - 2m_2 - 2v_1 - 2v_2)\Delta_{(a)_1(b)_2(a)_1(b)_2} + 2m_1\Delta_{(a)_1(b)_2(a)_2(b)_2} + 2m_2\Delta_{(a)_1(b)_1(a)_1(b)_2}$$

$$\dot{\Delta}_{(a)_2(b)_1(a)_2(b)_1} = \frac{1}{2N_1}\gamma_{(a)_2(a)_2} + \frac{1}{2N_2}\gamma_{(b)_1(b)_1} + (1 - \frac{1}{2N_1} - \frac{1}{2N_2} - 2m_1 - 2m_2 - 2v_1 - 2v_2)\Delta_{(a)_2(b)_1(a)_2(b)_1} + 2m_1\Delta_{(a)_2(b)_1(a)_2(b)_2} + 2m_2\Delta_{(a)_1(b)_1(a)_2(b)_1}$$

$$\dot{\Delta}_{(a)_1(b)_1(a)_2(b)_2} = \frac{1}{2N_1}\Gamma_{(ab)_1(a)_2(b)_2} + \frac{1}{2N_2}\Gamma_{(ab)_2(a)_1(b)_1} + (1 - \frac{1}{2N_1} - \frac{1}{2N_2} - 2m_1 - 2m_2 - 2v_1 - 2v_2)\Delta_{(a)_1(b)_1(a)_2(b)_2} + m_1\Delta_{(a)_1(b)_2(a)_2(b)_2}$$

$$+ m_1\Delta_{(a)_2(b)_1(a)_2(b)_2} + m_2\Delta_{(a)_1(b)_1(a)_2(b)_1} + m_2\Delta_{(a)_1(b)_1(a)_1(b)_2}$$

$$\bar{a}_{(a)_1(b)_2(a)_2(b)_2} = \frac{1}{2N_2}(r_{(a)_1(a)_2} + 2r_{(ab)_2(a)_1(b)_2}) + (1 - \frac{3}{2N_2} - m_1 - 3m_2 - 2v_1 - 2v_2)a_{(a)_1(b)_2(a)_2(b)_2} + m_1 a_{(a)_2(b)_2(a)_2(b)_2} + 2m_2 a_{(a)_1(b)_1(a)_2(b)_2}$$

$$+ m_2 a_{(a)_1(b)_2(a)_1(b)_2}$$

$$\bar{a}_{(a)_2(b)_1(a)_2(b)_2} = \frac{1}{2N_2}(r_{(b)_1(b)_2} + 2r_{(ab)_2(a)_2(b)_1}) + (1 - \frac{3}{2N_2} - m_1 - 3m_2 - 2v_1 - 2v_2)a_{(a)_2(b)_1(a)_2(b)_2} + m_1 a_{(a)_2(b)_2(a)_2(b)_2} + 2m_2 a_{(a)_1(b)_1(a)_2(b)_2}$$

$$+ m_2 a_{(a)_2(b)_1(a)_2(b)_1}$$

$$\bar{a}_{(a)_2(b)_2(a)_2(b)_2} = \frac{1}{2N_2}(r_{(a)_2(a)_2} + r_{(b)_2(b)_2} + 4r_{(ab)_2(a)_2(b)_2}) + (1 - \frac{6}{2N_2} - 4m_2 - 2v_1 - 2v_2)a_{(a)_2(b)_2(a)_2(b)_2} + 2m_2 a_{(a)_2(b)_1(a)_2(b)_2} + 2m_2 a_{(a)_1(b)_2(a)_2(b)_2}$$

# Appendix 4.

## Recursion Relationships for the Expected Values of the Identity Coefficients over Replicate Populations for the Variance of Homozygosity in a Single Population.

$$\phi'_{11} = (1-\mu)^2 \frac{1}{2N}\Big(1 + (2N-1)\phi_{11}\Big) +$$

$$2\mu(1-\mu)\left(\frac{2N-1}{2N}\right)\left(\frac{1}{k-1}\right)(1-\phi_{11}) +$$

$$O(\mu^2)$$

$$\Gamma'_{111} = (1-\mu)^3 \frac{1}{4N^2}\Big(1 + 3(2N-1)\phi_{11} + (2N-1)(2N-2)\Gamma_{111}\Big) +$$

$$3\mu(1-\mu)^2 \frac{1}{4N^2}\Big((2N-1)\left(\frac{1}{k-1}\right)(1-\phi_{11}) + (2N-1)(2N-2)\left(\frac{1}{k-1}\right)\Gamma_{11/1}\Big) +$$

$$O(\mu^2)$$

$$\Gamma'_{11/1} = (1-\mu)^3 \frac{1}{4N^2}\Big((2N-1)(1-\phi_{11}) + (2N-1)(2N-2)\Gamma_{11/1}\Big) +$$

$$2\mu(1-\mu)^2 \frac{1}{4N^2}\Big((2N-1)\left(\frac{1}{k-1}\right)(1-\phi_{11}) + (2N-1)(2N-2)\left(\frac{1}{k-1}\right)(\Gamma_{11/1} + \Gamma_{1/1/1})\Big) +$$

$$\mu(1-\mu)^2 \frac{1}{4N^2}\Big(1 + (2N-1)[3\phi_{11} + (1-\phi_{11})(1- \frac{1}{k-1})] + (2N-1)(2N-2)(\Gamma_{111} + (1- \frac{1}{k-1})\Gamma_{11/1})\Big) +$$

$$O(\mu^2)$$

$$\Gamma'_{1/1/1} = (1-\mu)^3 \frac{1}{4N^2}\Big((2N-1)(2N-2)\Gamma_{1/1/1}\Big) +$$

$$3\mu(1-\mu)^2 \frac{1}{4N^2}\Big(2(2N-1)(1- \frac{1}{k-1})(1-\phi_{11}) + (2N-1)(2N-2)[2(1- \frac{1}{k-1})\Gamma_{11/1} + (1- \frac{2}{k-1})\Gamma_{1/1/1}]\Big) +$$

$$O(\mu^2)$$

$$\Delta'_{1111} = (1-\mu)^4 \frac{1}{8N^3}\Big\{1 + 7(2N-1)\phi_{11} + 6(2N-1)(2N-2)\Gamma_{111} + (2N-1)(2N-2)(2N-3)\Delta_{1111}\Big\} +$$

$$4\mu(1-\mu)^3 \frac{1}{8N^3}\Big\{(2N-1)(\tfrac{1}{k-1})(1-\phi_{11}) + 3(2N-1)(2N-2)(\tfrac{1}{k-1})\Gamma_{11/1} + (2N-1)(2N-2)(2N-3)(\tfrac{1}{k-1})\Delta_{111/1}\Big\} +$$

$$O(\mu^2)$$

$$\Delta'_{111/1} = (1-\mu)^4 \frac{1}{8N^3}\Big\{(2N-1)(1-\phi_{11}) + 3(2N-1)(2N-2)\Gamma_{11/1} + (2N-1)(2N-2)(2N-3)\Delta_{111/1}\Big\} +$$

$$3\mu(1-\mu)^3 \frac{1}{8N^3}\Big\{(2N-1)(\tfrac{1}{k-1})(1-\phi_{11}) + (2N-1)(2N-2)(\tfrac{1}{k-1})(\Gamma_{41/1} + \Gamma_{1/1/1}) + (2N-1)(2N-2)(2N-3)(\tfrac{1}{k-1})(\Delta_{11/11} + \Delta_{11/1/1})\Big\} +$$

$$\mu(1-\mu)^3 \frac{1}{8N^3}\Big\{1 + (2N-1)[7\phi_{11} + (1- \tfrac{1}{k-1})(1-\phi_{11})] + 3(2N-1)(2N-2)(1- \tfrac{1}{k-1})\Gamma_{11/1} +$$

$$(2N-1)(2N-2)(2N-3)[(1- \tfrac{1}{k-1})\Delta_{111/1} + \Delta_{1111}]\Big\} +$$

$$O(\mu^2)$$

$$\Delta'_{11/11} = (1-\mu)^4 \frac{1}{8N^3}\Big\{(2N-1)(1-\phi_{11}) + 2(2N-1)(2N-2)\Gamma_{11/1} + (2N-1)(2N-2)(2N-3)\Delta_{11/11}\Big\} +$$

$$4\mu(1-\mu)^3 \frac{1}{8N^3}\Big\{(2N-1)(\tfrac{1}{k-1})(1-\phi_{11}) + (2N-1)(2N-2)(\tfrac{1}{k-1})(3\Gamma_{11/1} + \Gamma_{1/1/1}) + (2N-1)(2N-2)(2N-3)(\tfrac{1}{k-1})(\Delta_{11/1/1} + \Delta_{111/1})\Big\} +$$

$$O(\mu^2)$$

$$\Delta'_{11/1/1} = (1-\mu)^4 \frac{1}{8N^3}\Big\{(2N-1)(2N-2)\Gamma_{1/1/1} + (2N-1)(2N-2)(2N-3)\Delta_{11/1/1}\Big\} +$$

$$2\mu(1-\mu)^3 \frac{1}{8N^3}\Big\{2(2N-1)(2N-2)(\tfrac{1}{k-1})\Gamma_{1/1/1} + (2N-1)(2N-2)(2N-3)(\tfrac{1}{k-1})(2\Delta_{11/1/1} + \Delta_{1/1/1/1})\Big\} +$$

$$2\mu(1-\mu)^3 \frac{1}{8N^3}\Big\{(2N-1)(1- \tfrac{1}{k-1})(1-\phi_{11}) + (2N-1)(2N-2)[5(1- \tfrac{1}{k-1})\Gamma_{11/1} + (1- \tfrac{1}{k-1})\Gamma_{1/1/1}] +$$

$$(2N-1)(2N-2)(2N-3)[(1-\frac{1}{k-1})\Delta_{111/1} + (1-\frac{1}{k-1})\Delta_{11/11} + (1-\frac{2}{k-1})\Delta_{11/1/1}]\Big) +$$

$$O(\mu^{2})$$

$$\Delta'_{1/1/1/1} = (1-\mu)^{4}\frac{1}{8N^{3}}\Big((2N-1)(2N-2)(2N-3)\Delta_{1/1/1/1}\Big) +$$

$$4\mu(1-\mu)^{3}\frac{1}{8N^{3}}\Big(3(2N-1)(2N-2)(1-\frac{2}{k-1})\Gamma_{1/1/1} + (2N-1)(2N-2)(2N-3)[3(1-\frac{2}{k-1})\Delta_{11/1/1} + (1-\frac{3}{k-1})\Delta_{1/1/1/1}]\Big) +$$

$$O(\mu^{2})$$

# Appendix 5

## Derivation of the Recursion Relationship for Two-Locus, Fourth Order Moments

To determine the value of $\Phi'_{ijk/\ell mn/p/q}$ (the value of the coefficient in the next generation) requires that $i+j+k+\ell+m+n+p+q$ gametes be drawn at random, without replacement, from the present generation. For the moment, assume that none of these gametes are the result of mutation or recombination in the previous generation. Assume also, that no two of these gametes are copies of a single gamete in the previous generation. This will be true with approximate probability

$$1 - \nu_1(i+j+\ell+m+p+q) - \nu_2(i+k+\ell+n+p+q) - r(i+\ell+p+q)$$

$$- \frac{1}{2N} \tfrac{1}{2}(i+j+k+\ell+m+n+p+q)(i+j+k+\ell+m+n+p+q-1)$$

When none of these events occur, the sample of gametes will satisfy the required structure among alleles with probability

$$\Phi_{ijk/\ell mn/p/q}$$

This is the first term of the recursion relationship given below; no events have occured which change the probability from one generation to the next.

When two of the gametes are copies of one gamete in the previous generation it is necessary to determine their probability of identity / non-identity. It is convenient to consider the sample of gametes in groups (as in the definition of the coefficients). When two gametes from the same group are copies of one gamete, their probability of identity is one and therefore, the probability that the complete sample of gametes are identical / non-identical is equivalent to the probability for a sample with one less gamete in that group. For example, if two of the gametes in group I are copies of one gamete (this can happen in $\binom{i}{2}$ ways), then the probability is

$$\Phi_{i-1jk/\ell mn/p/q}$$

The probabilities are more complicated when two gametes from different groups are copies of one gamete in the previous generation. The derivation of these probabilities will be indicated with three examples. Example 1: From the definition, gametes in groups I and IV must have different alleles. Therefore if gametes from group I and group IV are copies of a single gamete in the previous generation, then the probability that these gametes are not identical is zero. Example 2: If gametes from group I and II are copies of one gamete (this can happen in $ij$ ways), then the probability is equivalent to the probability with one less gamete in group II because the identity of group I and II for this gamete is assured. ie:

$$\Phi_{ij-1k/\ell mn/p/q}$$

Example 3: If gametes from group II and III are copies of one gamete (this can happen in $jk$ ways), then the probability is equivalent to that with an extra gamete in group I, consisting of an A locus from group II linked to a B locus from group III. ie:

$$\Phi_{i+1j-1k-1/\ell mn/p/q}$$

The remaining probabilities are derived in a similar manner.

When recombination has occured, the resulting gamete is the union of two loci from two different gametes. Thus, if a gamete in group I is the result of recombination (this can happen in i ways with probability r), then the probability of identity / non-identity in the previous generation is

$$\Phi_{i-1j+1k+1/\ell mn/p/q}$$

That is, the probability with one less gamete in group I and one more in groups II and III. The probability of recombination can be ignored for gametes in groups II, III, V and VI since only a single locus is considered for these groups. For groups IV, VII and VIII the probability is similar to that for group I.

When a locus has a mutational event the probability of identity / non-identity is usually zero since an infinite alleles model is assumed. There are however, a few special cases where such mutations can contribute to the probability. These occur when there is only one allele of a locus (A or B) which must be different from one or more alleles of that locus. This allele will be different with probability one if a mutation occurs. It is then required that the remaining gametes have the correct probability structure. If more than one allele must be identical at the locus then a mutation can not be allowed. Therefore, to insure only one allele is present, a Dirac delta function is used

$$\delta(x) = 0 \quad \text{if} \quad x \neq 0$$

$$\delta(x) = 1 \quad \text{if} \quad x = 0$$

The mutational events which contribute to the probability are shown towards the end of the equation.

Putting all of this together, the general recursion relationship for the expected value of the identity coefficient over replicate populations is

$$\phi'_{ijk/\ell mn/p/q} = \Big( 1 - \frac{1}{2N}\tfrac{1}{2}(i+j+k+\ell+m+n+p+q)(i+j+k+\ell+m+n+p+q-1) - \nu_1(i+j+\ell+m+p+q)$$

$$- \nu_2(i+k+\ell+m+p+q) - r(i+\ell+p+q) \Big) \phi_{ijk/\ell mn/p/q} + \frac{1}{2N}\tfrac{1}{2}i(i-1)\phi_{i-1jk/\ell mn/p/q} +$$

$$\frac{1}{2N}ij\phi_{ij-1k/\ell mn/p/q} + \frac{1}{2N}ik\phi_{ijk-1/\ell mn/p/q} + \frac{1}{2N}\tfrac{1}{2}j(j-1)\phi_{ij-1k/\ell mn/p/q} +$$

$$\frac{1}{2N}jk\phi_{i+1j-1k-1/\ell mn/p/q} + \frac{1}{2N}jn\phi_{ij-1k/\ell mn-1/p+1/q} + \frac{1}{2N}jp\phi_{ij-1k/\ell mn/p/q} +$$

$$\frac{1}{2N}\tfrac{1}{2}k(k-1)\phi_{ijk-1/\ell mn/p/q} + \frac{1}{2N}km\phi_{ijk-1/\ell m-1n/p/q+1} + \frac{1}{2N}kq\phi_{ijk-1/\ell mn/p/q} +$$

$$\frac{1}{2N}\tfrac{1}{2}\ell(\ell-1)\phi_{ijk/\ell-1mn/p/q} + \frac{1}{2N}\ell m\phi_{ijk/\ell m-1n/p/q} + \frac{1}{2N}\ell n\phi_{ijk/\ell mn-1/p/q} +$$

$$\frac{1}{2N}\tfrac{1}{2}m(m-1)\phi_{ijk/\ell m-1n/p/q} + \frac{1}{2N}mn\phi_{ijk/\ell+1m-1n-1/p/q} + \frac{1}{2N}mq\phi_{ijk/\ell m-1n/p/q} +$$

$$\frac{1}{2N}\tfrac{1}{2}n(n-1)\phi_{ijk/\ell mn-1/p/q} + \frac{1}{2N}np\phi_{ijk/\ell mn-1/p/q} + \frac{1}{2N}\tfrac{1}{2}p(p-1)\phi_{ijk/\ell mn/p-1/q} +$$

$$\frac{1}{2N}\tfrac{1}{2}q(q-1)\phi_{ijk/\ell mn/p/q-1} + ir\phi_{i-1j+1k+1/\ell mn/p/q} + \ell r\phi_{ijk/\ell-1m+1n+1/p/q} +$$

$$pr\phi_{ij+1k/\ell mn+1/p-1/q} + qr\phi_{ijk+1/\ell m+1n/p/q-1} + \nu_1\delta(i-1)\delta(j+p)\phi_{00k+1/\ell mn/0/q} +$$

$$\nu_1\delta(j-1)\delta(i+p)\phi_{00k/\ell mn/0/q} + \nu_1\delta(p-1)\delta(i+j)\phi_{00k/\ell mn+1/0/q} + \nu_1\delta(\ell-1)\delta(m+q)\phi_{ijk/00n+1/p/0} +$$

$$\nu_1\delta(m-1)\delta(\ell+q)\phi_{ijk/00n/p/0} + \nu_1\delta(q-1)\delta(\ell+m)\phi_{ijk+1/00n/p/0} + \nu_2\delta(i-1)\delta(k+q)\phi_{0j+10/\ell mn/p/0} +$$

$$\nu_2\delta(k-1)\delta(i+q)\phi_{0j0/\ell mn/p/0} + \nu_2\delta(q-1)\delta(i+k)\phi_{0j0/\ell m+1n/p/0} + \nu_2\delta(\ell-1)\delta(n+p)\phi_{ijk/0m+10/0/q} +$$

$$\nu_2\delta(n-1)\delta(\ell+p)\phi_{ijk/0m0/0/q} + \nu_2\delta(p-1)\delta(\ell+n)\phi_{ij+1k/0m0/0/q}$$

The recursion relationship can be solved at equilibrium to give

$$\hat{\phi}_{ijk/\ell mn/p/q}\Big[(i+j+\ell+m+p+q)\theta_1 + (i+k+\ell+n+p+q)\theta_2 + (i+\ell+p+q)R +$$

$$(i+j+k+\ell+m+n+p+q)(i+j+k+\ell+m+n+p+q-1)\Big] = i(i-1)\hat{\phi}_{i-1jk/\ell mn/p/q} + 2ij\hat{\phi}_{ij-1k/\ell mn/p/q} +$$

$$2ik\hat{\phi}_{ijk-1/\ell mn/p/q} + j(j-1)\hat{\phi}_{ij-1k/\ell mn/p/q} + 2jk\hat{\phi}_{i+1j-1k-1/\ell mn/p/q} + 2jn\hat{\phi}_{ij-1k/\ell mn-1/p+1/q} +$$

$$2jp\hat{\phi}_{ij-1k/\ell mn/p/q} + k(k-1)\hat{\phi}_{ijk-1/\ell mn/p/q} + 2km\hat{\phi}_{ijk-1/\ell m-1n/p/q+1} + 2kq\hat{\phi}_{ijk-1/\ell mn/p/q} +$$

$$\ell(\ell-1)\hat{\phi}_{ijk/\ell-1mn/p/q} + 2\ell m\hat{\phi}_{ijk/\ell m-1n/p/q} + 2\ell n\hat{\phi}_{ijk/\ell mn-1/p/q} + m(m-1)\hat{\phi}_{ijk/\ell m-1n/p/q} +$$

$$2mn\hat{\phi}_{ijk/\ell+1m-1n-1/p/q} + 2mq\hat{\phi}_{ijk/\ell m-1n/p/q} + n(n-1)\hat{\phi}_{ijk/\ell mn-1/p/q} + 2np\hat{\phi}_{ijk/\ell mn-1/p/q} +$$

$$p(p-1)\hat{\phi}_{ijk/\ell mn/p-1/q} + q(q-1)\hat{\phi}_{ijk/\ell mn/p/q-1} + iR\hat{\phi}_{i-1j+1k+1/\ell mn/p/q} + \ell R\hat{\phi}_{ijk/\ell-1m+1n+1/p/q} +$$

$$pR\hat{\phi}_{ij+1k/\ell mn+1/p-1/q} + qR\hat{\phi}_{ijk+1/\ell m+1n/p/q-1} + \theta_1\delta(i-1)\delta(j+p)\hat{\phi}_{00k+1/\ell mn/0/q} +$$

$$\theta_1\delta(j-1)\delta(i+p)\hat{\phi}_{00k/\ell mn/0/q} + \theta_1\delta(p-1)\delta(i+j)\hat{\phi}_{00k/\ell mn+1/0/q} + \theta_1\delta(\ell-1)\delta(m+q)\hat{\phi}_{ijk/00n+1/p/0} +$$

$$\theta_1\delta(m-1)\delta(\ell+q)\hat{\phi}_{ijk/00n/p/0} + \theta_1\delta(q-1)\delta(\ell+m)\hat{\phi}_{ijk+1/00n/p/0} + \theta_2\delta(i-1)\delta(k+q)\hat{\phi}_{0j+10/\ell mn/p/0} +$$

$$\theta_2\delta(k-1)\delta(i+q)\hat{\phi}_{0j0/\ell mn/p/0} + \theta_2\delta(q-1)\delta(i+k)\hat{\phi}_{0j0/\ell m+1n/p/0} + \theta_2\delta(\ell-1)\delta(n+p)\hat{\phi}_{ijk/0m+10/0/q} +$$

$$\theta_2\delta(n-1)\delta(\ell+p)\hat{\phi}_{ijk/0m0/0/q} + \theta_2\delta(p-1)\delta(\ell+n)\hat{\phi}_{ij+1k/0m0/0/q}$$

where $\theta_1 = 4N\nu_1$, $\theta_2 = 4N\nu_2$, and $R = 4Nr$

# Appendix 6

## The Necessary Set of Equations for Two-Locus, Fourth Order Moments

Several properties of the coefficients must be used to insure the minimal number of equations. From the definition of the coefficients it is apparent that

$$\begin{aligned}\Phi_{ijk/\ell mn/p/q} &= \Phi_{\ell mn/ijk/q/p}\\ &= \Phi_{pjn/qmk/i/\ell}\\ &= \Phi_{qmk/pjn/\ell/i}\end{aligned}$$

A further group of identities among the coefficients occurs when a single A (or B) allele must be different from one or more A (or B) alleles. These probabilities can be obtained as the sum of two other probabilities. A well known example for a smngle locus is that the expected heterozygosity equals one minus the expected homozygosity. Similar reasoning leads to

$$\begin{aligned}
\Phi_{10k/\ell mn/0/q} &= \Phi_{00k+1/\ell mn/0/q} - \Phi_{00k/\ell mn/o/q+1}\\
\Phi_{01k/\ell mn/0/q} &= \Phi_{00k/\ell mn/0/q} - \Phi_{00k/\ell m+1n/0/q}\\
\Phi_{00k/\ell mn/1/q} &= \Phi_{00k/\ell mn+1/0/q} - \Phi_{00k/\ell+1mn/0/q}\\
\Phi_{1j0/\ell mn/p/0} &= \Phi_{0j+10/\ell mn/p/0} - \Phi_{0j0/\ell mn/p+1/0}\\
\Phi_{0j1/\ell mn/p/0} &= \Phi_{0j0/\ell mn/p/0} - \Phi_{0j0/\ell mn+1/p/0}\\
\Phi_{0j0/\ell mn/p/1} &= \Phi_{0j0/\ell m+1n/p/0} - \Phi_{0j0/\ell+1mn/p/0}\\
\Phi_{ijk/10n/p/0} &= \Phi_{ijk/00n+1/p/0} - \Phi_{ijk/00n/p+1/q}\\
\Phi_{ijk/01n/p/0} &= \Phi_{ijk/00n/p/0} - \Phi_{ij+1k/00n/p/0}\\
\Phi_{ijk/00n/p/1} &= \Phi_{ijk+1/00n/p/0} - \Phi_{i+1jk/00n/p/0}\\
\Phi_{ijk/1m0/0/q} &= \Phi_{ijk/0m+10/0/q} - \Phi_{ijk/0m0/0/q+1}
\end{aligned}$$

$$\Phi_{ijk/0m1/0/q} = \Phi_{ijk/0m0/0/q} - \Phi_{ijk+1/0m0/0/q}$$

$$\Phi_{ijk/0m0/1/q} = \Phi_{ij+1k/0m0/0/q} - \Phi_{i+1jk/0m0/0/q}$$

When a single A (or B) allele must be different from zero A (or B) alleles, then

$$\Phi_{01k/00n/0/0} = \Phi_{00k/01n/0/0} = \Phi_{00k/00n/0/0}$$

$$\Phi_{10k/00n/0/0} = \Phi_{00k/00n/0/1} = \Phi_{00k+1/00n/0/0}$$

$$\Phi_{00k/10n/0/0} = \Phi_{00k/00n/1/0} = \Phi_{00k/00n+1/0/0}$$

$$\Phi_{0j1/0m0/0/0} = \Phi_{0j0/0m1/0/0} = \Phi_{0j0/0m0/0/0}$$

$$\Phi_{1j0/0m0/0/0} = \Phi_{0j0/0m0/1/0} = \Phi_{0j+10/0m0/0/0}$$

$$\Phi_{0j0/1m0/0/0} = \Phi_{0j0/0m0/0/1} = \Phi_{0j0/0m+10/0/0}$$

These identities can be derived by expressing the coefficients as gene frequency moments.

Since it is assumed that $\nu_1 = \nu_2 = \nu$, at equilibrium the coefficients are symmetrical for the A and B loci and therefore

$$\begin{aligned}\Phi_{ijk/\ell mn/p/q} = \Phi_{ikj/\ell nm/q/p} &= \Phi_{\ell nm/ikj/p/q}\\ &= \Phi_{qkm/pnj/i/\ell}\\ &= \Phi_{pnj/qkm/\ell/i}\end{aligned}$$

The following definitions are used

if $i+j+k+\ell+m+n+p+q \leq 1$ then $\Phi_{ijk/\ell mn/p/q} = 1$

$$\Phi_{ijk/000/0/0} = \Phi_{ijk}$$

$$\theta = 4N\nu$$

$$R = 4Nr$$

The necessary systems of equations are derived from the general equilibrium equation given in Appendix 5 and using the above rules.

There are four independent equations and ten systems of equations with the remaining 46 coefficients. These are

$$\hat{\phi}_{020}(2+2\theta) = 2$$

$$\hat{\phi}_{030}(6+3\theta) = 6\hat{\phi}_{020}$$

$$\hat{\phi}_{040}(12+4\theta) = 12\hat{\phi}_{030}$$

$$\hat{\phi}_{020/020/0/0}(12+4\theta) = 4\hat{\phi}_{020} - 4\hat{\phi}_{030}$$

$$\begin{bmatrix} 2+4\theta+2R & -2R & 0 \\ -2 & 6+4\theta+R & -R \\ 0 & -8 & 12+4\theta \end{bmatrix} \ast \begin{bmatrix} \hat{\phi}_{200} \\ \hat{\phi}_{111} \\ \hat{\phi}_{022} \end{bmatrix} = \begin{bmatrix} 2 \\ 4\hat{\phi}_{020} \\ 4\hat{\phi}_{020} \end{bmatrix}$$

$$\begin{bmatrix} 6+5\theta+2R & -2R & 0 \\ -4 & 12+5\theta+R & -R \\ 0 & -12 & 20+5\theta \end{bmatrix} \ast \begin{bmatrix} \hat{\phi}_{210} \\ \hat{\phi}_{121} \\ \hat{\phi}_{032} \end{bmatrix} = \begin{bmatrix} 2\hat{\phi}_{020} + 4\hat{\phi}_{200} \\ 2\hat{\phi}_{030} + 6\hat{\phi}_{111} \\ 2\hat{\phi}_{030} + 6\hat{\phi}_{022} \end{bmatrix}$$

$$\begin{bmatrix} 12+6\theta+2R & -2R & 0 \\ -6 & 20+6\theta+R & -R \\ 0 & -16 & 30+6\theta \end{bmatrix} \ast \begin{bmatrix} \hat{\phi}_{220} \\ \hat{\phi}_{131} \\ \hat{\phi}_{042} \end{bmatrix} = \begin{bmatrix} 2\hat{\phi}_{030} + 10\hat{\phi}_{210} \\ 2\hat{\phi}_{040} + 12\hat{\phi}_{121} \\ 2\hat{\phi}_{040} + 12\hat{\phi}_{032} \end{bmatrix}$$

$$\begin{bmatrix} 6+6\theta+3R & -3R & 0 & 0 \\ -2 & 12+6\theta+2R & -2R & 0 \\ 0 & -8 & 20+6\theta+R & -R \\ 0 & 0 & -18 & 30+6\theta \end{bmatrix} \ast \begin{bmatrix} \hat{\phi}_{300} \\ \hat{\phi}_{211} \\ \hat{\phi}_{122} \\ \hat{\phi}_{033} \end{bmatrix} = \begin{bmatrix} 6\hat{\phi}_{200} \\ 2\hat{\phi}_{111} + 8\hat{\phi}_{210} \\ 12\hat{\phi}_{121} \\ 12\hat{\phi}_{032} \end{bmatrix}$$

$$\begin{bmatrix} 12+7\theta+3R & -3R & 0 & 0 \\ -4 & 20+7\theta+2R & -2R & 0 \\ 0 & -12 & 30+7\theta+R & -R \\ 0 & 0 & -24 & 42+7\theta \end{bmatrix} \ast \begin{bmatrix} \hat{\phi}_{310} \\ \hat{\phi}_{221} \\ \hat{\phi}_{132} \\ \hat{\phi}_{043} \end{bmatrix} = \begin{bmatrix} 6\hat{\phi}_{210} + 6\hat{\phi}_{300} \\ 2\hat{\phi}_{121} + 4\hat{\phi}_{220} + 10\hat{\phi}_{211} \\ 6\hat{\phi}_{131} + 12\hat{\phi}_{122} \\ 6\hat{\phi}_{042} + 12\hat{\phi}_{033} \end{bmatrix}$$

$$\begin{bmatrix} 12+8\theta+4R & -4R & 0 & 0 & 0 \\ -2 & 20+8\theta+3R & -3R & 0 & 0 \\ 0 & -8 & 30+8\theta+2R & -2R & 0 \\ 0 & 0 & -18 & 42+8\theta+R & -R \\ 0 & 0 & 0 & -32 & 56+8\theta \end{bmatrix} \ast \begin{bmatrix} \hat{\phi}_{400} \\ \hat{\phi}_{311} \\ \hat{\phi}_{222} \\ \hat{\phi}_{133} \\ \hat{\phi}_{044} \end{bmatrix} = \begin{bmatrix} 12\hat{\phi}_{300} \\ 6\hat{\phi}_{211} + 12\hat{\phi}_{310} \\ 2\hat{\phi}_{122} + 20\hat{\phi}_{221} \\ 24\hat{\phi}_{132} \\ 24\hat{\phi}_{043} \end{bmatrix}$$

$$\begin{bmatrix} 12+6\theta+2R & 0 & -2R & 0 \\ 0 & 12+6\theta+2R & -2R & 0 \\ -2 & -4 & 20+6\theta+R & -R \\ 0 & 0 & -16 & 30+6\theta \end{bmatrix} \times \begin{bmatrix} \hat{\Phi}_{200/020/0/0} \\ \hat{\Phi}_{110/010/0/1} \\ \hat{\Phi}_{111/020/0/0} \\ \hat{\Phi}_{022/020/0/0} \end{bmatrix} = \begin{bmatrix} 2\hat{\Phi}_{020} - 2\hat{\Phi}_{030} + 2\hat{\Phi}_{200} - 2\hat{\Phi}_{210} \\ 4\hat{\Phi}_{111} - 2\hat{\Phi}_{210} \\ 2\hat{\Phi}_{020/020/0/0} + 2\hat{\Phi}_{111} + 2\hat{\Phi}_{022} - 4\hat{\Phi}_{121} \\ 2\hat{\Phi}_{020/020/0/0} + 4\hat{\Phi}_{022} - 4\hat{\Phi}_{032} \end{bmatrix}$$

$$\begin{bmatrix} 12+7\theta+3R & -2R & -R & 0 & 0 \\ -4 & 20+7\theta+2R & 0 & -2R & 0 \\ -4 & 0 & 20+7\theta+2R & -2R & 0 \\ 0 & -8 & -4 & 30+7\theta+R & -R \\ 0 & 0 & 0 & -24 & 42+7\theta \end{bmatrix} \times \begin{bmatrix} \hat{\Phi}_{200/010/0/1} \\ \hat{\Phi}_{111/010/0/1} \\ \hat{\Phi}_{201/020/0/0} \\ \hat{\Phi}_{112/020/0/0} \\ \hat{\Phi}_{023/020/0/0} \end{bmatrix} = \begin{bmatrix} 2\hat{\Phi}_{111} - 2\hat{\Phi}_{300} \\ 4\hat{\Phi}_{121} - 4\hat{\Phi}_{211} + 4\hat{\Phi}_{110/010/0/1} \\ 2\hat{\Phi}_{022} + 2\hat{\Phi}_{210} - 2\hat{\Phi}_{121} - 2\hat{\Phi}_{211} + 4\hat{\Phi}_{200/020/0/0} \\ 2\hat{\Phi}_{121} + 2\hat{\Phi}_{032} - 4\hat{\Phi}_{122} + 6\hat{\Phi}_{111/020/0/0} \\ 4\hat{\Phi}_{032} - 4\hat{\Phi}_{033} + 6\hat{\Phi}_{022/020/0/0} \end{bmatrix}$$

$$\begin{bmatrix} 12+8\theta+4R & -4R & 0 & 0 & 0 & 0 \\ -2 & 20+8\theta+3R & -2R & -R & 0 & 0 \\ 0 & -8 & 30+8\theta+2R & 0 & -2R & 0 \\ 0 & -8 & 0 & 30+8\theta+2R & -2R & 0 \\ 0 & 0 & -12 & -6 & 42+8\theta+R & -R \\ 0 & 0 & 0 & 0 & -32 & 56+8\theta \end{bmatrix} \times \begin{bmatrix} \hat{\Phi}_{200/000/0/2} \\ \hat{\Phi}_{201/010/0/1} \\ \hat{\Phi}_{112/010/0/1} \\ \hat{\Phi}_{202/020/0/0} \\ \hat{\Phi}_{113/020/0/0} \\ \hat{\Phi}_{024/020/0/0} \end{bmatrix} = \begin{bmatrix} 4\hat{\Phi}_{210} - 4\hat{\Phi}_{300} \\ 2\hat{\Phi}_{121} + 2\hat{\Phi}_{220} - 2\hat{\Phi}_{211} - 2\hat{\Phi}_{310} + 6\hat{\Phi}_{200/010/0/1} \\ 4\hat{\Phi}_{131} - 4\hat{\Phi}_{221} + 10\hat{\Phi}_{111/010/0/1} \\ 2\hat{\Phi}_{032} + 2\hat{\Phi}_{220} - 2\hat{\Phi}_{122} - 2\hat{\Phi}_{221} + 10\hat{\Phi}_{201/020/0/0} \\ 2\hat{\Phi}_{131} + 2\hat{\Phi}_{042} - 4\hat{\Phi}_{132} + 12\hat{\Phi}_{112/020/0/0} \\ 4\hat{\Phi}_{042} - 4\hat{\Phi}_{043} + 12\hat{\Phi}_{023/020/0/0} \end{bmatrix}$$

$$\begin{bmatrix} 12+8\Theta+4R & -4R & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -2 & 20+8\Theta+3R & -2R & -R & 0 & 0 & 0 & 0 & 0 \\ 0 & -4 & 30+8\Theta+2R & 0 & -4 & -2R & 0 & 0 & 0 \\ 0 & -8 & 0 & 30+8\Theta+2R & 0 & -2R & 0 & 0 & 0 \\ 0 & 0 & -R & 0 & 20+8\Theta+3R & 0 & -2 & -2R & 0 \\ 0 & 0 & -8 & -2 & 0 & 42+8\Theta+R & 0 & -8 & -R \\ 0 & 0 & 0 & 0 & -4R & 0 & 12+8\Theta+4R & 0 & 0 \\ 0 & 0 & 0 & 0 & -8 & -2R & 0 & 30+8\Theta+2R & 0 \\ 0 & 0 & 0 & 0 & 0 & -32 & 0 & 0 & 56+8\Theta \end{bmatrix} \times \begin{bmatrix} \dot{\Phi}_{200/200/0/0} \\ \dot{\Phi}_{200/111/0/0} \\ \dot{\Phi}_{111/111/0/0} \\ \dot{\Phi}_{200/022/0/0} \\ \dot{\Phi}_{101/110/1/0} \\ \dot{\Phi}_{111/022/0/0} \\ \dot{\Phi}_{100/100/1/1} \\ \dot{\Phi}_{101/021/1/0} \\ \dot{\Phi}_{022/022/0/0} \end{bmatrix} =$$

$$\begin{bmatrix} 4\dot{\Phi}_{200} - 8\dot{\Phi}_{210} + 4\dot{\Phi}_{300} \\ 2\dot{\Phi}_{111} - 4\dot{\Phi}_{121} + 2\dot{\Phi}_{211} + 4\dot{\Phi}_{200/020/0/0} - 4\dot{\Phi}_{200/010/0/1} \\ 8\dot{\Phi}_{111/020/0/0} - 8\dot{\Phi}_{111/010/0/1} \\ 2\dot{\Phi}_{022} - 4\dot{\Phi}_{032} + 2\dot{\Phi}_{122} + 4\dot{\Phi}_{200/020/0/0} - 4\dot{\Phi}_{201/020/0/0} \\ 4\dot{\Phi}_{110/010/0/1} - 4\dot{\Phi}_{200/010/0/1} \\ 4\dot{\Phi}_{022/020/0/0} - 8\dot{\Phi}_{112/020/0/0} + 4\dot{\Phi}_{111/020/0/0} \\ 0 \\ 4\dot{\Phi}_{111/020/0/0} - 4\dot{\Phi}_{201/020/0/0} + 2\dot{\Phi}_{110/010/0/1} - 2\dot{\Phi}_{111/010/0/1} \\ 8\dot{\Phi}_{022/020/0/0} - 8\dot{\Phi}_{023/020/0/0} \end{bmatrix}$$

## Appendix 7

## Taylor's Series Approximation to $r^2$:

## An Example

Let $D = f_{11} - pq$

| | Population | | | Average |
|---|---|---|---|---|
| | #1 - #10 | #11 | #12 | |
| $f_{11}$ | 0.9998 | 0.9700 | 0.9700 | - |
| $f_{12}$ | 0.0001 | 0.0100 | 0.0100 | - |
| $f_{21}$ | 0.0001 | 0.0200 | 0.0100 | - |
| $f_{22}$ | 0.0000 | 0.0000 | 0.0100 | - |
| $D^2$ | $1.0000\times10^{-16}$ | $4.0000\times10^{-8}$ | $9.2160\times10^{-5}$ | $7.6833\times10^{-6}$ |
| $p(1-p)q(1-q)$ | $9.9980\times10^{-9}$ | $1.9404\times10^{-4}$ | $3.8416\times10^{-4}$ | $4.8192\times10^{-5}$ |
| $D^2p(1-p)q(1-q)$ | $9.9980\times10^{-25}$ | $7.7616\times10^{-12}$ | $3.5404\times10^{-8}$ | $2.9510\times10^{-9}$ |
| $p^2(1-p)^2q^2(1-q)^2$ | $9.9960\times10^{-17}$ | $3.7652\times10^{-8}$ | $1.4758\times10^{-7}$ | $1.5436\times10^{-8}$ |
| $D^2/p(1-p)q(1-q)$ | $1.0002\times10^{-8}$ | $2.0614\times10^{-4}$ | $2.3990\times10^{-1}$ | $2.0009\times10^{-2}$ |

Using the averages as expected values, the Taylor's series approximations to $r^2$ are

$$r^2 = E[D^2/p(1-p)q(1-q)]$$

$$= 0.020009$$

$$r^2 = E[D^2]/E[p(1-p)q(1-q)]$$

$$= 0.15943$$

$$r^2 = \frac{E[D^2]}{E[p(1-p)q(1-q)]}\left(1 - \frac{E[D^2 p(1-p)q(1-q)]}{E[D^2]E[p(1-p)q(1-q)]} + \frac{E[p^2(1-p)^2 q^2(1-q)^2]}{E^2[p(1-p)q(1-q)]}\right)$$

$$= -0.051557$$