THE INDIRECT ESTIMATION OF MUTATION RATES IN MAN

A thesis

submitted for the degree of

Doctor of Philosophy

in the

Australian National University

by

KULDEEP KUMAR BHATIA
John Curtin School of Medical Research
The Australian National University
Canberra

JULY, 1981

STATEMENT

This thesis embodies the results of studies conducted in the Department of Human Biology, John Curtin School of Medical Research, Australian National University, Canberra under the supervision of Dr R.L. Kirk, during the tenure of an Australian National University Scholarship.

The analyses presented are, unless otherwise indicated, entirely my own work.

KULDEEP KUMAR BHATIA

ACKNOWLEDGEMENTS

During the preparation of this volume I have received enormous help from my Guru Dr R.L. Kirk. I owe a deep debt of gratitude to him for his personal interest, encouragement and supervision throughout the course of my studies. I am also grateful to:

Dr N.M. Blake for extensive help in the compilation of data and also for making available results of his unpublished laboratory data.

Dr S.W. Serjeantson and Dr S.R. Wilson for their advice on the analysis of the data.

Professor Trefor Jenkins for providing his unpublished results on Kgalgadi and discussions on the genetic polymorphisms in S. African populations.

Dr Ranajit Chakraborty for providing unpublished manuscripts and helpful comments on data analysis.

Dr J.V. Neel for comments on some of the papers included here.

Dr Y.S. Teng for teaching some methods of protein separation and Drs Graham Jones and Philip Board for fruitful discussions.

I have also received enormous help from Ms Janis McGettrick, Christine Hayes and Anne Thompson in understanding the experimental aspects of the data.

My thanks are due to my wife Saroj Bhatia for help in the preparation of the manuscript.

Mrs Robbie Williams typed the manuscript most carefully. My sincere thanks.

This project was made possible by the joint award of a Ph.D. Scholarship by the Government of India and the Australian National University. The leave without salary granted by Panjab University, Chandigarh allowed me to pursue my research interests.

Abstract

The need for generating reliable estimates of mutation rate in man has been emphasized by a number of geneticists. The present study estimates the rate of mutation at cistron level using data on electromorphs in a number of human populations.

The number and frequency of private electromorphs, both rare and polymorphic, have been enumerated in relatively isolated populations from Australia, Papua New Guinea, India and sub-Saharan Africa. It is noticed that the recovery of these electrophoretic variants is directly related to the number of genes sampled, the size of the polypeptide subunit electrophoresed and the molecular constraints arising from the assembly of multimeric proteins.

Mutation rates in 38 individual populations have been generated in the present study by five different methods. The average estimates of mutation rate obtained by the methods of Kimura and Ohta (1969) and Rothman and Adams (1978) are 5.99×10^{-6} and 6.39×10^{-6}/locus per generation respectively. The estimates by the rare allele methods of Nei (1977) and Chakraborty (1981) and by a new method, suggested in the present study, utilizing singletons, are comparatively smaller, being 4.62×10^{-6}, 2.55×10^{-6} and 2.86×10^{-6}/

locus per generation respectively.

The estimates of mutation rate generated here show regional/ethnic differences. The significance of these differences, however, cannot be properly evaluated since large standard errors are known to be associated with the estimates of the number of alleles segregating and the population size, two of the parameters indispensible to the indirect approaches.

The suggestion of Nei *et al.*(1976b) regarding the variability in mutation rates of various protein loci has been tested using the data on rare alleles in 12 different human populations. The results indicate the existence of significant correlation between subunit molecular weights and the number and frequency of rare alleles.

The present results have indicated the existence of variability in the estimates of mutation rate more or less at the same level as that generated from indirect estimates of "classical" traits. The need for deciding the order of magnitude of mutation rate in man is alive as ever.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES Page

Chapter 1

HISTORICAL PERSPECTIVE

The average rate at which human genes mutate is a parameter not only of considerable importance in evolutionary genetics, but also has certain immediate practical implications because of the contribution mutation makes to human ill-health (Neel, 1978a; Knudson, 1979). However, despite the very considerable effort devoted to the subject and the imposing body of knowledge which has accummulated during the past eight decades, our understanding of the spontaneous and induced mutation rates in higher organisms has not been proportional to the recent advances in other fields of genetical research (Neel, 1977).

One of the major reasons for this lack of knowledge is that the subject of mutation rates in eukaryotes has been discussed generally on the basis of visible Mendelian traits. The estimates generated from these traits suffer from two obvious sources of bias: (a) the relationship of the observed changes to the alterations in the genetic code is unknown, and (b) the loci selected are generally those at which mutations are already known to have occurred and their inclusion in the sample of loci studied is a function of their mutation rate (Cavalli-Sforza and

Bodmer, 1971; Yasuda, 1973).

The recent developments in the fields of protein separation and the demonstrated co-linearity of a gene and its polypeptide product have now made it possible to relate the changes at the polypeptide or aminoacid level to the coding nucleotides. In addition, the advances in the fields of histochemical staining and two dimensional electrophoresis have reduced considerably the reliance on traits with demonstrated mutability. Besides, the processes of mutagenesis and DNA repair are better understood and this may be used to explain much of the variability in the mutation rates encountered over various loci.

Meanwhile, the mathematical theory of population genetics has become much more sophisticated. Particularly noteworthy is the theoretical framework provided by the manipulation of differential equations (Kimura, 1964) and the branching process models (Rothman and Adams, 1978; Thompson and Neel, 1978). While the diffusion models allow one to describe the behaviour of mutant alleles by considering the random changes resulting from random sampling of gametes during reproduction, as well as the deterministic changes caused by mutation and selection (Kimura, 1979a) the branching process models given by Neel and Thompson (1978) and Rothman and Adams (1978) relate the probability distributions of current observations to generate

a likelihood function which permits a direct classical approach to the estimation of mutation rates. In addition, models to explain genetic variations at electromorph (Ohta and Kimura, 1973) allelic (Kimura and Crow, 1964) and nucleotide (Kimura, 1969) level are useful in relating the mutation rates to the corresponding observations. The sampling theory of neutral alleles (Ewens, 1972) and algorithms to relate the sample estimates to the population values (Nei, 1977; Rothman and Adams, 1978; Chakraborty, 1981) have also proved useful in the newer approaches to mutation rate estimation.

Although it would be ideal to generate mutation rates using direct observations of fresh mutations in a population, the magnitude of the associated errors and the prohibitive cost of monitoring (Neel *et al.*, 1980a) make such a program a difficult, although not impossible proposition. Evidence from other sources, however, lends itself to statistical manipulation to generate indirectly estimates of mutation rates which may be reliable.

In the present study I use such indirect methods to generate estimates of average mutation rate from electrophoretic data in various human populations. In addition, I relate the mutation rates for individual polypeptide chains to the physico-chemical and configurational constraints of the molecule. Previous estimates of mutation

rates then will be discussed in the light of new evidence.

1.1. The Nature of Mutation.

According to prevailing concepts mutation results from molecular "slips" during DNA replication and repair and also during meiotic processes. Meiosis is implicated because higher mutation rates occur during meiosis than mitosis (Magni and von Borstel, 1962). Numerous human diseases exhibit alterations in the mechanisms by which the damaged DNA is repaired and mutations reproduced. Investigations of these diseases, notably xeroderma pigmentosum (XP), ataxia telangiectasia (AT) and Fanconi's anaemia (FP) have shown that human repair systems are complex inter-related systems with distinct features (Cleaver, 1978). The paternal age effects and the sex differences in sex-linked mutations seen in achondroplasia and Lesch-Nyhan disease respectively support the role of replication in mutation process (Vogel and Motulsky, 1979), although the adequacy of the basic models of DNA replication and base mispairing is being increasingly doubted (Drake, 1978).

The present explanations of the mutagenic processes, however, have evolved from a number of chemical and physical hypotheses put forward in the last sixty years. Bateson (1928) conceived of mutation as a presence-absence situation

according to which all mutations are due to the loss of the normal gene. The theory explained most of the phenomena prevailing at that time (Auerbach, 1971) and is applicable, even now, in part to null mutations of proteins, but certainly does not explain most of the mutational changes being unfolded recently.

Muller (1922) conceived of a mutation as some kind of 'autocatalytic' change, i.e. the basic function of the gene is retained although its form is altered, a description which was eventually supported by Watson and Crick's (1953) model of mutation through DNA replication and base mis-pairing processes.

The generally accepted view of mutagenesis that the variability caused by induced mutations is in no way different from that produced by spontaneous mutations, has led to a number of explanations attributable to different mutagenic involvements of irradiation and exposure to chemicals. To begin with, there was not much similarity in the approach of those investigators studying the mutagenic effects of radiation and those working on chemically induced mutations and the field was dominated by biophysical questions, particularly the question of validity of the target theory and the role of the indirect effect of radiation (Lea, 1946). The situation with chemical

mutagens was clarified by the analysis of the nature of mutations induced by simple chemicals in both plants (Gierer and Mundry, 1958) and bacterial viruses (Freese, 1959). The eras of radiation genetics and chemical mutagenesis, have however, only touched a part of the mutational spectrum. It may be relevant to point out here that the high exposure to mutagenic chemicals is only a recent phenomenon in human populations and that the naturally occuring ionizing radiation is far too weak to account for the rates at which spontaneous mutation rates have prevailed in the past. In fact, the role of metabolic processes in the induction and maintenance of an optimal level of spontaneous mutations cannot be ignored (Auerbach, 1978).

This question of the insufficiency of physico-chemical explanation of the mutational events has seen another dimension in the recently proposed mutational theories of carcinogenesis (Knudson, 1971). The carcinogenic evidence of mutation, in turn, should help unravel the nature of mutations in the near future.

The elucidation of the genetic code (for a review, see Jukes 1978) together with advances in working out the primary structure of protein molecules has meant that the precise change of the aminoacid sequence in a mutant protein can

often be identified (Brock, 1978). The bulk of information comes from the study of mutant haemoglobins (Masters and Holmes, 1975). This genetically determined variation in the protein structure, however, takes a variety of forms whose end effects vary from lethal to benign. Most of these may be attributable to structural gene mutations (Harris, 1981). The variable expression of human genes in the synthesis of proteins may, however, be partly attributed to regulatory gene mutations (Kazazian *et al*, 1977). This loss of activity in proteins resulting from mutations leads generally to ill effects, although the evidence is accummulating that the loss of protein activity may not be only restricted to harmful states (Neel, 1978a).

Many studies of protein variation have employed methods of protein separation, usually by electrophoresis or have depended on the presence of a gross difference in enzyme activity (Nyhan, 1977). Some have explored differences in the kinetic properties of enzymes. More recently, investigators have begun to apply immunochemical methods to decipher the type of mutation (Ben-Yoseph *et al*, 1978).

1.2. The Estimation Procedures.

The first procedure to estimate mutation rate indirectly, even before direct estimations were

made, was given by Danforth (1921). Since then a number of indirect procedures to estimate mutation rate using mutation-selection equilibrium (Haldane, 1935), average life span of a mutant prior to extinction (Ewens, 1964; Kimura and Ohta, 1969), lethal and detrimental equivalents (Morton, 1960), sex-ratio changes (Traut, 1969) and number of different alleles (Nei, 1977; Rothman and Adams, 1978; Chakraborty, 1981) have been proposed. In addition, a number of direct, semi-direct (Morton, 1959) and relative (Neel, 1977; Zouros, 1979) methods to estimate mutation rate have been suggested.

1.2.1. Indirect methods.

Danforth (1921) suggested that the average number of generations a mutant survives in a population ($\bar{t}_o$) is a function of its selective value. The mutation rate for slightly unfavourable deleterious mutations μ, can be estimated as:

$$\mu = x/\bar{t}_o \qquad (1.1)$$

where x is the frequency of the trait in the population. This argument, however, overlooked the fact that irrespective of the selective disadvantages, most of the new mutations are eventually lost to the population (Fisher, 1930). Besides, the rarity of information on $\bar{t}_o$ has made the use of this approach limited.

Using the diffusion approximations, Ewens (1964) has given the average time until extinction under the infinite alleles model as

$$\bar{t}_o = 2 \int_0^1 x^{-1}(1-x)^{\theta-1}\, dx \qquad (1.2)$$

where $\theta=4N\mu$ is the scaled rate of mutation. Another simple expression for $\bar{t}_o$ is given by Kimura and Ohta (1969) as

$$\bar{t}_o = 2\left(\frac{N_{ev}}{N}\right)\log 2N \qquad (1.3)$$

where N_{ev} and N are respectively the variance effective and actual size of the population. Although both the models use diffusion approximations to arrive at the expressions for estimating $\bar{t}_o$, the former model incorporates the rate of mutation whereas the latter model is essentially a two allele model giving a conditional mean extinction time. Li and Neel (1974) and Li (1978), however, find the formula (1.3) unrealistic for the population structures actually obtained and have shown through simulation studies that the Kimura-Ohta equation leads to overestimation. Note, however, that the argument has come a long way from Danforth's formula.

Haldane (1935) used the mutation-selection equilibrium to estimate mutation rate, a method which involved a balance sheet of loss and renewal. The rate of renewal, sufficient to balance the loss, was called mutation rate. Haldane's equilibrium

equation is given as

$$2N\mu = KN(1-f)x \qquad (1.4)$$

Where N is the population size, x is the frequency of the trait in question, f its genetic fitness or fecundity and K is a constant depending upon the loss of the number of genes through the death of an affected individual. The value of K is one for autosomal dominants, two for autosomal recessives and 2/3 for sex-linked recessive traits. Neel (1962) modified the formulation for autosomal recessive characters to include the role of consanguinity in exposing hidden traits, as:

$$\mu = (1-f)[\alpha x+(1-\alpha)x^2] \qquad (1.5)$$

where μ is the expected amount of the inbreeding coefficient and K=1. Nei and Imaizumi (1963) modified the equation to estimate mutation rate for rare recessive traits.

The errors involved in estimating mutation rates using classical Mendelian traits have been listed extensively. For direct observations these include paternity errors, phenocopies and polygenic inheritance. For indirect procedures, in addition to the errors listed for the direct methods, the validity of the assumption of genetic equilibrium is the most questionable. In addition, the reliability of the estimates of fertility differentials and for the recessive traits, the role of

inbreeding make the utility of these estimates rather doubtful. An excellent review of these probable errors is given by Neel (1962).

With the advent of molecular genetics, a point mutation was recognized as an alteration in the DNA by the substitution of a purine or pyrimidine base. This can result in a large number of mutations for each polypeptide and this, in turn, led to a number of attributes of genetic variation, in addition to the frequency of the allelic trait in question, to be used for estimating the mutation rate.

Some of these parameters used for estimating the rate of mutation are: the expected number of different alleles, i.e. total, rare or single copy alleles given under the infinite alleles model (Karlin and McGregor, 1967; Ewens, 1972; Nei, 1977, Chakraborty, 1981 and Rothman and Adams, 1978), the expected number of electromorphs given under the stepwise mutation model (Kimura and Ohta 1975), the total number of nucleotide sites segregating under the infinite sites model (Kimura and Ohta, 1969), heterozygosity under all the three models of mutation namely infinite alleles (Kimura and Crow, 1964), stepwise (Kimura and Ohta, 1975) and infinite sites model (Kimura, 1969) and the probability of monomorphism, P_m under the infinite alleles model (Kimura and Ohta, 1971). For strict neutral cases the estimation of $\theta(=4N\mu)$ from heterozygosity is known to yield a biased estimate of μ

under the infinite alleles model (Ewens and Gillespie, 1974). Ewens (1972), on the other hand, has demonstrated the number of different alleles to be a sufficient statistic for θ under the infinite alleles model. A number of different sampling procedures to estimate the mutation rate from the number of different alleles recovered in a sample are given by Ewens (1972), Ewens (1974), Watterson (1974), Nei (1977), Rothman and Adams (1978) and Chakraborty (1981). The equations used in the various estimation procedures and their relative efficiencies are discussed in a later section.

The concept of recurrent mutation is, however, presented in a different form when the estimates $\hat{K}$ and $\hat{F}$ are based on electromorphs. This is because of the redundancy in the genetic code and of cases where there is no charge change due to the substitution of an amino acid with one of equal charge. Besides, a large number of charge changes may coalesce to form an electromorph (Nei and Chakraborty, 1976; Chakraborty and Nei, 1976). The situation may, however, be resolved to a certain extent by the used of a number of other electrophoretic conditions.

1.2.2. Direct methods.

The direct method for dominant traits was first used by Gunther and Penrose (1935) and consists of simply counting all sporadic cases

(those with normal parents and negative family histories) of the trait. A factor of 0.5 is used to make adjustment for the fact that each individual possesses two genes at each locus. Harris *et al.* (1974) and Neel *et al.* (1980a) have used negative results in population surveys to give the maximum limit of the mutation rate by using the formula

$$p = (1-\mu)^{2n} \qquad (1.6)$$

where p is the probability of detectingno fresh mutation in a sample of 2n genes.

1.2.3. Semi-direct methods.

Semi-direct estimates of mutation rate are obtained if the proportion of fresh mutations in the population is estimated, rather than identified (as is the case with direct methods). Using the data on sporadic cases (not familial or chance isolated cases) the rate of mutation (μ) is calculated as

$$\mu = \hat{\theta}'I/\text{locus per generation} \qquad (1.7)$$

where $\hat{\theta}'$ is the estimated proportion of sporadic cases among all affected and I is the incidence of trait in the general population.

Dewey, Barrai, Morton and Mi (1965) have suggested two alternative procedures for the estimation of θ'. For non-recessive sporadic cases they suggest the use of segregation analysis given by Morton (1959) to obtain the maximum

likelihood estimates of θ'. For recessive characters the consanguinity analysis of Chung, Robison and Morton (1959) is used by Dewey *et al.* (1965) to estimate the proportion of sporadic cases (θ').

The incidence of the trait in the general population (I) is obtained by a direct enumeration or can also be estimated indirectly by using segregation as well as consanguinity analysis. For data obtained through incomplete selection, Barrai *et al.* (1965) define the incidence of the trait as:

$$I = A/\Pi N$$

where A is the number of probands in the populations at a given time, N is the size of the population and Π is the probability of ascertainment (the probability that an affected individual is a proband).

In traits with heterogeneous aetiology, i.e. the traits which are produced by more than one locus and also sometimes include some non-genetic factors, equation (1.7) overstimates μ. Dewey *et al.* (1965) have suggested the use of the detrimental equivalents approach of Morton (1960) to estimate the number of contributory loci (ℓ) for these traits. Accordingly if X is the contribution of all factors other than autosomal

recessives then the incidence of cases from normal parents I is given as:

$$I = 1-\Sigma c_i e^{-(A+B\alpha_i)}$$

$$\simeq A+B\alpha \qquad (1.9)$$

where A, B and α are panmictic load, inbreeding load and average inbreeding coefficient respectively, given as:

$$A = X+\Sigma q^2 t,$$

$$B = \Sigma q(1-q)t, \text{ and}$$

$$\alpha = \Sigma c_i \alpha_i$$

in which q is the recessive gene frequency, t is the penetrance and the summation for the values of A and B is over ℓ loci. The average coefficient of inbreeding (α) is obtained by summation over all the couples. The coefficient of inbreeding (α_i) and c_i is the frequency of couples with the coefficient of inbreeding α_i. If Q is the mean frequency of the trait per contributory locus, then A and B are given as:

$$A \simeq \ell Q^2+X, \text{ and} \qquad (1.10a)$$

$$B \simeq \ell Q(1-Q) \qquad (1.10b)$$

where ℓ is the number of contributory loci. From equations (1.10a) and (1.10b), the value of ℓ is obtained as:

$$\ell > B^2/A \qquad (1.11)$$

Thus:-

$$\mu = \frac{\theta' I}{\ell} \text{ per locus per generation} \qquad (1.12)$$

A similar approach can be used for lethal characters using the lethal equivalents method of Morton *et al.* (1956).

Despite its intuitive appeal the semidirect method using segregation analysis suffers from ascertainment problems. Another limitation of segregation analysis is its restriction to single sibships, although Elston and Stewart (1971) have suggested procedures involving pedigree analysis. However, the correction for ascertainment effects in pedigree analysis is still not adequate (Elston, 1973; Elston and Yelverton, 1975). The usefulness of this approach may increase in the near future with newer procedures to distinguish homozygotes from heterozygotes.

The utility of lethal and detrimental equivalent methods has also suffered from the inadequacy of the procedures to provide clear-cut answers. As pointed out by Schull and Neel (1965), a large proportion of the genetic load ascertained in the populations is segregational rather than mutational, which makes the interpretation of the lethal equivalents difficult. On the other hand, assumption of the uniformity of the mutation rates over all the loci contributing to the detrimental equivalents, leads to underestimation of the number of loci or, in turn, overestimation of the

average estimate of the mutation rate per locus (Cavalli-Sforza and Bodmer, 1971). Some of the problems involved have been rectified by Morton *et al.* (1977) in their re-analysis of the Colchester data on mental retardation.

Translating the load into an actual mutation rate, however, is an uncertain proposition because of the different contributions made by megaphenic and microphenic characters. No estimates of the two separate contributions have been made so far in man (Mukai, 1979).

Using the isolation by distance models, Morton *et al.* (1973) has suggested another approach to the question of mutation rate estimations. According to them, if Q is the mean frequency of the trait per locus and m the systematic pressure (Morton, 1977), then μ is given as

$$\mu = mQ/\text{locus/generation} \qquad (1.13)$$

1.2.4. Relative methods.

The question of relative magnitudes of mutation rates between different populations, loci, ages and sexes, is basic to the question of variability in mutation rates. The estimates of the relative proportions have the added advantage that one can extrapolate the basic mutation rate for simpler traits to traits with difficult aetiology. In addition, the relative estimates can provide some

insight into the mechanisms of mutagenesis.

An aspect of the relative estimates of mutation rate which generated a good deal of controversy is the sex-difference in the mutation rate. If a mutation is considered as a replication error, the chances are that the mutation rates per locus per generation will be larger in males than in the females due to the differences in the number of cell divisions during gametogenesis. Argued similarly, one can associate chronological age with mutation rate.

It was shown by Haldane (1947) that in an infinite population at equilibrium for a sex-linked lethal gene one-third of all the male cases will be children of the non-carrier mothers. Any deviation from this ratio of 1:2 may be assigned to the difference in the relative mutation rates in males (v) and females (u).

The ratio is given as:

$$\frac{v}{u} = \frac{s}{\theta'} - 2 \qquad (1.14)$$

where s is the selection coefficient against the hemizygote male and θ' is the proportion of sporadic cases among all affected males. If v = u and s = 1, then $\theta' = 1/3$. If I is the incidence of the trait then the mutation rate in females is:

$$u = \theta' I/\text{locus/generation} \qquad (1.15)$$

and the value of v may be estimated using equations (1.14) and (1.15) as:

$$v^2 = (s-2\theta')I \qquad (1.16)$$

Davie and Emery (1978) have suggested another procedure to estimate the relative rates using the sex-ratio of the children of the carrier and normal mother. It is, however, a less efficient statistical method and is prone to the changes in the ratios if reproductive compensation is involved, as suggested by Lange *et al.* (1978).

Neel (1977) used the data for the number of rare variants per 1000 persons to estimate relative mutation rates between homologous loci and also between populations. As well he suggested the use of Ewens' (1972) test statistics to find the significance of these differences.

Zouros (1979) has suggested a least square method for the estimation of relative magnitudes of mutation rate between any two loci. Using the relationship between the expected amount of homozygosity, E(F) and θ (= $4N\mu$) given by Kimura and Crow (1964) for the infinite alleles model, the relative rate of mutation between loci x and y in the ith population is given by him as:

$$\frac{\hat{\theta}_{Fix}}{\hat{\theta}_{Fiy}} = \frac{\hat{F}_{ix} - 1}{\hat{F}_{iy} - 1} \qquad (1.17)$$

An estimate of F is obtained as:

$$\hat{F}_{ix} = \Sigma P_{ixm}^2 \qquad (1.18)$$

Where P_{ixm} is the gene frequency of the mth allele at the xth locus in ith population. For the two loci, x and y, studied in r populations there will be r estimates of μ_x/μ_y. Plotting $\hat{\theta}_{Fix}$ *versus* $\hat{\theta}_{Fiy}$ for i = 1 to r gives r points on a two-dimensional space. The slope of the straight line which passes through the origin and minimizes the summation of squares of distances of r points from this line then gives an estimate of μ_x/μ_y. Over ℓ loci the relative rates of mutation are obtained by scaling $\Sigma\mu_x$ to 1. One can similarly use the relationship between expected amount of homozygosity and mutation rate under the step-wise model of mutation (Zouros, 1979).

Similar estimation procedures are obtained if the data on observed number of alleles (k) in a sample of 2n genes are used for estimating θ. Using the equation of Ewens (1972) for relating E(k) to θ under the infinite alleles model Zouros (1979) gives the relative rate of mutation as:

$$\frac{k_{ix}}{k_{iy}} = \frac{\hat{\theta}_{kix} \sum_{j=0}^{2n-1} (\hat{\theta}_{kix} + j)^{-1}}{\hat{\theta}_{kiy} \sum_{j=0}^{2n-1} (\hat{\theta}_{kiy} + j)^{-1}} \qquad (1.19)$$

where k_{ix} and k_{iy} are the observed number of alleles in the populations x and y respectively and 2n is the number of genes sampled in populations x and y respectively.

The superiority of $\hat{\theta}_k$ over $\hat{\theta}_F$ has been shown already by Ewens and Gillespie (1974) for strict neutrality. According to them $\hat{\theta}_F$ overestimates θ by 40% or more. Although this bias is claimed to be rendered negligible by Nei (1975) and Li (1979) if a number of loci are used to estimate F, this bias is increased greatly (Ewens, 1979). Since one locus is used at one time to estimate $\hat{F}$ in Zouros' method, this bias in the estimation of θ is unavoidable.

However, it is difficult to extend Zouros' method to human populations because of large variations in the effective population sizes, which are considered to be of similar size in his method. However, considering that θ is linearly related to expected homozygosity (Zouros, 1979), the relative proportions of heterozygosity contributed by rare alleles can be used to estimate the relative rates of mutation over different loci, which does not take into consideration the size of populations. This simplified version given by Bhatia (1981) is, however, intuitive rather than based on formal theoretical basis.

1.2.5. Poissonian approximation methods.

An indirect method to estimate mutation rates for both germinal (μ_g) and the somatic (μ_s) mutations has been given by Hethcote and Knudson (1978) for two-event (or more) mutational processes. The quantitative model using the Poisson distribution relates the age-specific incidence of the character explicitly to the number of divisions of embryonal cells and to the rates of somatic mutations per cell division. The Poissonian approximations have, however, been challenged (Matsunaga, 1978) on the ground that both penetrance and expressivity in the gene carrier can be defined as a variable determined by genetic and environmental factors and not by a Poisson distribution of tumors formed.

1.3. Estimates of Mutation Rates.

Before the developments of molecular genetics, geneticists had estimated that the rate of spontaneous mutation per locus is of the order of 10^{-5} per generation in many higher organisms such as fruitfly, corn and man (Nei, 1975). A number of researchers, however, believed that this rate is too high (Cavalli-Sforza and Bodmer, 1971; Yasuda, 1973). Serious sources of error in such estimates included the occurrence of phenocopies, incomplete penetrance, the polygenic nature of some of the mutations and the bias in the loci sampled.

1.3.1. Indirect estimates.

In the indirect methods more reliable estimates can be obtained for autosomal dominants and sex-linked recessive characters than for autosomal recessives (Neel, 1962).

The first indirect estimates of mutation rate were for syndactyly and polydactyly and were given by Danforth (1921) by noting that the incidence of each trait is less than 1 in 2,000 genomes and that each trait persists, on an average, three generations, which gives a mutation frequency of less than one in 6,000 genomes. Modern estimates began with Haldane (1935) and Gunther and Penrose (1935) who reported mutation rates for haemophilia and epiloia as 2×10^{-5} and 8×10^{-6} per locus per

generation respectively.

Lists of mutation rates for various traits have been compiled by Crow (1961), Stevenson and Kerr (1967), Conneally (1974), Edwards (1974) and Vogel and Rothenberg (1975). Only a few limited estimates of mutation rates for classical traits have been added or revised since then.

For a sample of 49 recessive loci on the X-chromosome compiled by Stevenson and Kerr (1967), a mean mutation rate of $1.97 \pm 0.76 \times 10^{-6}$ is given by Yasuda (1973). It is worth noting that the majority of these 49 loci exhibit mutation rates of less than 1×10^{-6} per locus per generation. Recent refinements in the laboratory analyses of carrier detection have allowed a comparison of the results generated by using Haldane's (1935) equilibrium method with the direct investigations. For example, Gardner-Medwin (1970), in a systematic survey, has found that the mutation rate for Duchenne muscular dystrophy diagnosed by CPK tests (10.5×10^{-5}) is in close agreement with the estimate obtained by using Haldane's equilibrium method. On the other hand, Lesch-Nyhan disease does not exhibitequality in mutation rates between the two sexes (Franke *et al.*, 1976; 1977) which reflects on the utility of equilibrium models. The need for using different equilibrium situations for males and females is obvious.

Yasuda (1973) has also given the mean values for autosomal dominants and recessives as 2.50± 0.61 x 10^{-5} and 2.93± 0.45 x 10^{-5} for samples of 23 and 9 loci respectively. These mean values, however, differ from the values for sex-linked traits by an order of magnitude. This difference may be due to a lack of data on traits with low mutation rates for autosomal dominants and recessives (Yasuda, 1973).

Indirect estimates of mutation rate from neutral allele models have also been made in several studies. In isolated South American Indians, Neel (1973) suggested a mutation rate of 6-8 x 10^{-5} per locus.

Later, with a somewhat different approach, a value of 4.8 x 10^{-5} per locus was obtained (Neel and Rothman, 1978). A series of other estimates on different world populations have been given by Nei (1977), Tchen *et al.* (1978), Chakraborty and Roychoudhury (1978), Bhatia *et al.* (1979, 1981), Bhatia (1981b) and Chakraborty (1981). The range of these estimates will be discussed in a later section.

1.3.2. Direct estimates.

In addition to the direct estimates of mutation rate for autosomal dominant traits discussed in an

earlier section, there are only a few reports on mutation rates derived directly from electromorph data. Kimura and Ohta (1973) extrapolated the proportion of fresh mutations in the haemoglobin variants to the frequency of variants in a Japanese survey. Their calculations yield a mutation rate of 3.3×10^{-5} per cistron per generation. Dubinin and Altukhov (1979) reported a mutation rate of $\simeq 6 \times 10^{-5}$ per cistron per generation in a population from USSR for a set of protein loci. In a preliminary report Neel *et al.* (1980b) have also reported the recovery of one probable mutation in the offspring of "proximally exposed" parents from Hiroshima and Nagasaki. However, these authors concluded the data insufficient to provide a worthwhile estimate of mutation rate.

Harris *et al.*(1974), however, found no mutations in 113,478 locus tests on inhabitants of the United Kingdom. Neel *et al.* (1980a) has reported no recovery of fresh mutations in 94,796 locus tests on Amerindians of central and South America and 105,649 locus tests in Ann Arbor, Michigan. 208,196 locus tests on Japanese (Neel *et al.* 1980b) also did not reveal any fresh mutations. Neel *et al.* (1980a) give the upper limit of the mutation rate at 95% confidence level for 522,119 locus tests as 0.6×10^{-5}/locus

per generation in the combined total sample of all these populations.

1.3.3. Semi-direct estimates.

One of the recent estimates of mutation rate based on the detrimental equivalents is given by Morton *et al.* (1977). By discriminating between the socio-familial and the biological types of mental retardation, they have estimated an average per locus mutation rate of 2.4×10^{-5} (for at least 351 loci with mutation rate per gamete of 0.008). These results contrast with an earlier estimate of 1.32×10^{-5} per locus (0.0019 per gamete for 144 mimic genes) given by Dewey *et al.* (1965). It is instructive to note that Cavalli-Sforza and Bodmer (1971) had already corrected the minimum number of loci for mental retardation to 338, a value quite similar to the one given later by Morton *et al.* (1977), after a revision using more sophisticated methods of segregation analysis. It will be interesting to see how much correction is to be made to a similar estimate on deaf mutism by Dewey *et al.* (1965).

Gardner-Medwin (1970) and Yasuda and Kondo (1980) have used the semi-direct method to estimate mutation rates in Duchenne muscular dystrophy. The two estimates are 10.5×10^{-5} and 6.3×10^{-5} per locus per generation respectively. Bucher *et al.*

(1980) have used various methods to estimate the proportion of sporadic cases for Duchenne muscular dystrophy, although no exact estimate of the mutation rate was given.

1.3.4. Relative estimates.

The relative estimates of mutation rate for both sexes for sex-linked recessive characters have generated a good deal of controversy. Franke *et al.* (1976) and Winter (1980) have estimated an approximate ratio of 10:1 in male/female mutation rates. This ratio is, however, not significantly different from one. A similar magnitude is seen in haemophilia, although Duchenne muscular dystrophy does not reveal any such difference (Yasuda and Kondo, 1980; Morton, 1979).

Neel (1977) found that the rate of mutation of the structural genes for polypeptides of haemoglobins and carbonic anhydrase did not show significant difference whereas the variants of PGM_1 exhibited differences in the relative rates of mutation. Zouros (1979) has reported large ratios between the mutation rates for various polypeptide chains in *Drosophila* although Bhatia (1981a) found the mutation rates in man to show a much smaller inter-cistron range.

1.3.5. Poissonian approximation estimates.

Hethcote and Knudson (1978) have provided two estimates of mutation rates for somatic cells

by the Poissonian approximation methods as 3.9×10^{-7} and 4.8×10^{-7} mutations per locus per cell division. The order of magnitude for the mutation rates given by Hethcote and Knudson (1978) is quite close to other such estimates based on direct observations on HLA variants in cultivated human lymphoid cells as also in other somatic cells, *in vitro* as well as *in vivo*, by Pious and Soderland (1977), Stamatoyannopoulos (1979), Stamatoyannopoulos *et al.* (1980) and van Zeeland and Simons (1976).

Chapter 2

THEORETICAL FORMULATION

Most of the indirect approaches for the estimation of mutation rate outlined in Chapter 1 were formulated before the models of selective equivalence (Kimura and Crow, 1964; Kimura, 1968; Ohta, 1975; 1976; Li, 1979a; Kimura, 1979b and Ewens and Li, 1980) were put forward. A number of variations of these models of mutation, based on an appreciation of the laboratory techniques employed to detect the genetic variability, have since been presented under the general framework of neutral mutation theory.

2.1. Neutrality Models.

The neutrality models will be discussed below under three headings:

1. Infinite alleles model
2. Infinite sites model
3. Step-wise mutation model

2.1.1. Infinite alleles model.

Kimura and Crow (1964) formulated the infinite alleles model which assumes that an infinite sequence of A_1, A_2,---, alleles can occur at any particular locus. These alleles are selectively equivalent and any gene mutates with fixed but unknown probability to give rise to an allele of an entirely new type not currently or previously seen in the population. The

model predicts that the variability within a species in terms of average heterozygosity $\bar{H}$ per gene will be determined essentially by the product of the effective population size N_e and the mutation rate μ per generation, rather than by N_e and μ separately (Kimura, 1979a).

2.1.2. Infinite sites model.

To describe the genetic heterogeneity at the smallest level, Kimura (1969) suggested a model of mutation for the total genome at the level of nucleotides. The theory is also applicable to a small group of nucleotides, say a codon. Since the number of available sites (nucleotides or codons) for mutation is sufficiently large while the mutation rate per site is very low, every fresh mutation occurs at a site at which no mutant forms are segregating already. This assumption is known as infinite sites model.

The model of infinite sites was actually given by Kimura (1969) for the whole genome. However, to a sufficient degree of approximation the model is applicable to a gene locus or cistron which is made up of a finite number (several hundreds) of nucleotides or codons, provided the number of segregating sites per cistron is low (Kimura, 1979b). In this regard, the infinite sites model and infinite alleles model are identical provided there is no intra-cistronic recombination.

2.1.3. Step-wise mutation model.

Both the above models, however, describe incompletely the genetic variability demonstrated by standard electrophoretic procedures. While the infinite sites model does not account for the conformational changes in the protein molecules, the infinite alleles model does not take into account the various charge changes which coalesce to form a single electromorph. Ohta and Kimura (1973) suggested a new model named the step-wise mutation model. Other workers have termed this the ladder rung model, the charge state model and the electrophoretic model to relate the demonstrated electrophoretic variability to the basis of mutation. A mutation leading to a charge change gives rise to a step forward or backward on the electrophoretic screen.

2.2. Generalised Neutrality.

Ohta (1973) modified the neutral mutation random drift hypothesis of Kimura (1968) and King and Jukes (1969) to incorporate selective constraint (negative selection). The model is based on the idea that selective neutrality is the limit when the selective disadvantages become infinitely small (Kimura and Ohta, 1974). Ohta (1975, 1976) described in detail the role of deleterious mutations in maintaining polymorphisms

and later (Ohta, 1977) investigated a model in which selection coefficients against mutants follow an exponential distribution. According to Kimura (1979b) the model of Ohta (1977) has a drawback in that it cannot accommodate enough mutations which behave effectively as neutral when the population size gets large. Kimura (1979b) suggested that the generalized neutrality model should incorporate selection coefficients which follow a gamma distribution.

This argument has been developed further by Watterson (1977, 1978a), Li (1977, 1978), Ewens (1979a) and Ewens and Li (1980). Li (1979a) has divided the mutation rate into two separate parts i.e. μ_{total} and $\mu_{neutral}$. If f_o is the fraction of selectively neutral mutations in the total mutations then

$$\mu_{neutral} = f_o \, \mu_{total}$$

In the limit ($f_o \to 1$),

$$\mu_{neutral} = \mu_{total}$$ (Kimura,1977).

2.3. Neutrality-Theory and Estimation of Mutation Rate.

The question whether data on electrophoretic variants in natural populations are in accord with the null hypothesis of strict neutrality has been investigated in several studies. In these a number of parameters of genetic variation, especially the

amount of heterozygosity, the number of different alleles and the number of rare alleles, have been used by Ewens (1977, 1979b), Ewens and Feldman (1976), Watterson (1977, 1978a, b), Watterson and Anderson (1980), Nei *et al.* (1976a), Fuerst *et al.* (1977), Chakraborty *et al.* (1978, 1980).

Since the utility of these various parameters in arguing the cause of neutrality is well established, these same parameters can be usefully employed as statistics to estimate the rate of mutation as suggested by Neel (1973), Nei (1977), Neel and Thompson (1978), Rothman and Adams (1978) and Chakraborty (1981). A number of other relationships between the mutation rate and different aspects of genetic data given by Ewens (1964, 1972, 1979a), Kimura and Crow (1964), Karlin and McGregor (1967) and Watterson (1974) can also be used to generate estimates of mutation rate.

2.4. Mathematical Formulation.

In the following sections various mathematical formulations, their simple approximate and exact forms, as used by these various workers, will be outlined. These different methods utilize a variety of data and use quite different modelling procedures. Broadly, we can group these approaches under three headings, i.e.

1. Diffusion approximations
2. Branching process models
3. Equilibrium models

2.4.1 Diffusion approximation methods.

Let $\Phi(x)$ define the frequency spectrum having the property that $\Phi(x)dx$ is the mean number of alleles in the population with frequency in $(x,x+dx)$. According to the infinite alleles model of Kimura and Crow (1964), the expression is given as

$$\Phi(x) = \theta\, x^{-1}(1 - x)^{\theta-1} \qquad (2.1)$$

where $\theta = 4N\mu$, in which N is the actual size of the population and μ is the rate of mutation of the allelic level.

For infinite sites model, the frequency spectrum, $\Phi(x)$, is given by the irreversible mutation model of Wright (1938) as

$$\Phi_1(x) = 4N\nu x^{-1} \qquad (2.1a)$$

where ν is the rate of mutation per codon or nucleotide.

For the step-wise mutation model, Kimura and Ohta (1975) have given the equation for this frequency spectrum as

$$\Phi(x) = \frac{\Gamma(\theta+B'+1)}{\Gamma(\theta)\Gamma(B'+1)}\,(1-x)^{\theta-1}x^{B'-1} \qquad (2.2)$$

where $\theta = 4N\mu$, $B' = (1+4N\mu - \sqrt{1+8N\mu})/(\sqrt{1+8N\mu}-1)$ and $\Gamma(.)$ is a gamma function. Hereafter, we shall denote $\Phi(x)$ under the infinite alleles, infinite sites and step-wise mutation models as $\Phi(x)dx$, $\Phi_1(x)dx$ and $\Phi_2(x)dx$, respectively. Nei *et al* (1976b) give the frequency spectrum for the variable mutation rates model with the gamma distribution, as

$$\Phi(x) = \frac{\bar{\theta}x^{-1}(1-x)^{-1}}{[1-\bar{\theta}/\alpha]\log(1-x)^{\alpha+1}} \qquad (2.3)$$

where $\alpha = \bar{\theta}^2/V_\theta$ is the parameter of the gamma distribution in which $\bar{\theta}$ and V_θ are the mean and variance of the variate in question (θ), respectively. When the variance of θ approaches zero with θ constant, α tends to ∞. In this case (2.3) tends to (2.1), as expected.

The frequency spectrum $\Phi(x)$ can be used to find expressions for two different parameters namely the mean number of different alleles in the population ($\bar{K}$) and the probability that any two genes are of different allelic types ($\bar{H}$).

These parameters are given as

$$\bar{K} = \int_{\frac{1}{2N}}^{1} \Phi(x)dx \qquad (2.4)$$

$$\bar{H} = \int_{0}^{1} x(1-x)\Phi(x)dx \qquad (2.5)$$

2.4.1.1. The sampling theory.

Suppose a sample of n individuals (or 2n genes) is drawn from a population of size N (or 2N genes). It is assumed that $n \ll N$ so that, although sampling is without replacement, binomial sampling formulae can be used to a sufficient degree of approximation.

Let $\sum_{j=1}^{2n} k_j = k$ represent the number of alleles in the sample, where j represents the number of copies by which an allele is represented in the sample and k_j is the random number of alleles represented by j genes. Let n_1 represent the number of genes of the first allelic type, n_2 the number of genes of the second allelic type, and so on. The summation $\sum_{j=1}^{k} n_j$ equals the total number of genes samples or 2n. Ewens (1972) and Karlin and McGregor (1972) have shown that the probability of the random vector

$(k; n_1, n_2, \ldots, n_k)$ is,

$$\Pr(k; n_1, n_2, \ldots, n_k) = \frac{2n!\,\theta^k \Gamma(\theta)}{k!\,n_1\, n_2 \ldots n_k\, \Gamma(2n+\theta)} \qquad (2.7)$$

From (2.7) the probability distribution of k is found as

$$\Pr(k) = S_{2n}^{k}\, \theta^k\, \Gamma(\theta)/\Gamma(2n+\theta) \qquad (2.8)$$

where S^{k}_{2n} is the absolute value of a Stirling number of the first kind (Ewens, 1972). From (2.7) and (2.8) we get the conditional distribution of allelic frequency as

$$\Pr(n_1, n_2, \ldots, n_k | k) = \frac{2n!}{k! S^{k}_{2n} n_1 n_2 \ldots n_k} \qquad (2.9)$$

Note that this conditional distribution is independent of θ which implies that k is a sufficient statistic for θ. By sufficient statistic, we mean that all the information about θ is contained in k and that the relative frequencies of various allelic types do not yield any additional information on θ. Statistical theory shows that the inclusion of $n_1, n_2 \ldots, n_k$ or any statistic derived thereof in the inferential procedure may only introduce noise. The maximum likelihood estimate $\hat{\theta}_k$ of θ, given k, is then found by using the equation (Ewens, 1972),

$$E(k) = \theta \int_0^1 [1-(1-x)^{2n}] x^{-1} (1-x)^{\theta-1} \, dx \qquad (2.10)$$

$$= \theta \sum_{j=0}^{2n-1} (\theta+j)^{-1}$$

Similarly, once k is observed, we estimate θ by $\hat{\theta}$, given as the solution of the equation

$$k = \frac{\hat{\theta}_k}{\hat{\theta}_k} + \frac{\hat{\theta}_k}{\hat{\theta}_k+1} + \ldots \frac{\hat{\theta}_k}{\hat{\theta}_k+m-1}$$

The value of $\hat{\theta}_k$ can be calculated numerically for any given value of k and n. Ewens (1972) has provided ready reference tables which can be used to estimate θ_k.

The sampling equation (2.10) can also be used to obtain $E(k;q_1,q_2)$, the expected number of alleles in the sample whose sample frequency is between q_1 & q_2. This is given as

$$E(k;q_1,q_2)=\theta\int_{q_1}^{q_2}[1-(1-x)^{2n}]x^{-1}(1-x)^{\theta-1}dx \qquad (2.12)$$

An approximate solution of (2.12) is

$$E(k) \simeq \theta\log(q_2/q_1) - \theta(\theta-1)[q_2-q_1] \qquad (2.14)$$

We define k^* as the random number of segregated sites. The $E(k^*)$ is given as

$$E(k^*) = \theta_1\int_{\frac{1}{2n}}^{1} x^{-1}dx$$

$$= \theta\log(2n) \qquad (2.15)$$

Where $E(k^*)$ is the expected number of segregating sites and $\theta_1 = 4N\nu$.

Another estimator of θ which has been commonly used is $\hat{\theta}_F$ based on the amount of homozygosity ($\hat{F}$) obtained in the sample. Kimura and Crow (1964) have calculated the mean value of $\hat{F}$ to be

$$E(\hat{F}) = (2N(1-\mu)^2-2N+1)^{-1}$$

$$\simeq (1+\theta)^{-1} \qquad (2.16)$$

when $\hat{F}$ is defined as

$$\hat{F} = \Sigma(2n_j)^2/2n^2$$

The estimator $\hat{\theta}_F$ is then given as

$$\hat{\theta}_F \nearrow \hat{F}^{-1} - 1 \qquad (2.18)$$

This estimator is, however, biased. Ewens and Gillespie (1974) have shown through simulation that the mean value of $\hat{\theta}_F$ is consistently about 40% or more in excess than the actual value of θ. Besides these estimates have poor sampling properties (Bodmer and Cavalli-Sforza, 1972).

A number of variations of the random variable k, the number of different alleles, have been suggested for estimating θ. These include:

1. The total number of alleles (k)
2. Rare allelic variants (k_r)
3. Singletons (k_s)

The use of the total number of alleles, k, to estimate θ_k using infinite alleles neutrality model of Kimura and Crow (1964) also runs into difficulty when the data are on electromorphs. This is because the more frequent electromorphs encompass a variable number of silent allelic substitutions for which information is lost in the estimation procedures. Besides, the role of selection in maintaining these high frequency alleles cannot be ascertained.

Nei (1977) and Chakraborty (1981) have advocated the use of alleles segregating in the lower frequency ranges, specially those with sample frequencies of less than 0.01 or 0.05 for calculating $\hat{\theta}_k$. Although the choice of these arbitrarily designated rare alleles for

calculating $\hat{\theta}_k$ entails loss of a certain amount of information, this loss is compensated by a better fit of the data on electromorphs to the infinite alleles model when only rare alleles are used.

2.4.1.1.1. Total number of alleles (k)

In the infinite alleles model, the estimate $\hat{\theta}_k$ of θ is found from the equations

$$k = 1 + \frac{\hat{\theta}_k}{\hat{\theta}_k+1} + \frac{\hat{\theta}_k}{\hat{\theta}_k+2} + \ldots + \frac{\hat{\theta}_k}{\hat{\theta}_k+n-1}$$

values of $\hat{\theta}$ for given n and k can be found either from the tables in Ewens (1972) or numerically.

For the infinite sites model, Ewens (1974) gives the estimator $\hat{\theta}^*_k$, as

$$\hat{\theta}^*_k = k^*/S_{2n-1} \qquad (2.24)$$

where S_{2n-1} is defined as $\sum_{j=1}^{2n-1} j^{-1}$ and for large 2n is approximated to $\log_e(2n-1)+\gamma$.

2.4.1.1.2. Number of rare alleles (k_r).

Nei (1977) and Chakraborty (1981) have advocated the use of number of rare alleles, i.e. only those alleles whose sample frequency is equal to or smaller than a specified value of q (taken arbitrarily as 0.01 or 0.05) to estimate θ. Their argument is based on the realization that the probability of a low frequency electromorph being composed of more than one aminoacid substitution is very low. In addition, the role of selection in maintaining these rare alleles in the

population is negligible.

Nei (1977) equated the mean number of alleles segregating in the population in the frequency range $[(2n)^{-1},q]$ to the number of alleles, whose sample frequency is within the range, $[(2n)^{-1},q)$ recovered in a sample (k_r) using the steady state formula of Wright (1938). This method of moments approach yields an estimator of $\hat{\theta}_{k_r}$ defined by

$$k_r = \int_{2n^{-1}}^{q} \Phi_1(x)dx$$

$$= \hat{\theta}\log(2nq) \qquad (2.25)$$

Using Nei's approach, the number of rare alleles in the sample are equated to the mean number through the equations

$$k_r = \int_{2n^{-1}}^{q} \Phi(x)dx$$

$$\simeq \hat{\theta}_{k_r}\log(2nq) - \hat{\theta}_{k_r}(\hat{\theta}_{k_r}-1)\{q-\tfrac{1}{2n}\} \qquad (2.26)$$

to give $\hat{\theta}_{k_r}$ where $\Phi(x)$ is already defined. Equations (2.25) and (2.26) yield almost similar values for large 2n.

Chakraborty (1981) has extended the equation (2.19) given by Chakraborty *et al* (1980) to rare alleles to provide an estimator of θ_{k_r}, which is given as

$$k_r = \sum_{j=1}^{[2nq]} \frac{\hat{\theta}_{k_r}}{j} \frac{(2n)!}{(2n-j)!} \frac{\Gamma(2n+\hat{\theta}_{k_r-j})}{\Gamma(2n+\hat{\theta}_{k_r})} \qquad (2.29)$$

where [2nq] is the largest integrator in inner expression and $\Gamma(.)$ is the gamma function. The approximate

solutions of (2.29) is given as

$$k_r = A\hat{\theta}_{k_r} - B\hat{\theta}^2_{k_r} \qquad (2.31)$$

where

$$A = \sum_{j=1}^{[2nq]} j^{-1} \quad \text{and} \quad B = \sum_{j=1}^{[2nq]} (2n-j)^{-1}$$

For the infinite sites model, the binomial sampling equations lead to

$$E(k_r^*) = \sum_{j=1}^{[2nq]} \binom{2n}{j} \int_0^1 x^j (1-x)^{2n-j} \Phi_1(x)dx$$

$$= \theta \sum_{j=1}^{[2nq]} \binom{2n}{j} \int_0^1 x^{j-1}(1-x)^{2n-j}dx$$

$$= \theta \sum_{j=1}^{[2nq]} \binom{2n}{j} \beta(j,2n-j+1) \qquad (2.36)$$

where $\beta(.,.)$ is the beta function. This leads to

$$k_r^* = \hat{\theta} \sum_{j=1}^{[2nq]} j^{-1} = \theta A \qquad (2.37)$$

which for large [2nq] is given as

$$k_r^* = \hat{\theta}(\log 2nq + \gamma) \qquad (2.38)$$

Notice that (2.25) and (2.38) are quite close for large [2nq], although (2.37) always yields smaller estimates of $\hat{\theta}_{k_r}$.

Before turning to consider the step-wise mutation

model, it is important to notice that the infinite sites model without recombination is identical to the infinite alleles model. Besides, for alleles with low frequency the expected number of different alleles in a sample approximates well the number of different sites segregating (Nei, 1977). Any difference in the results using these two models must, therefore, be related to different aspects of allelic data used as input.

For the step-wise mutation model, the form of equation (2.29) is given as

$$E(k_r) = \sum_{j=1}^{[2nq]} \binom{2n}{j} B(B'+j;\theta+2n-j)/B(\theta,B'+1) \qquad (2.39)$$

where $B(.,.)$ is the beta function and B' is as given in (2.2). This expression is practically identical to (2.29) for $\theta<0.01$ as seen in extensive numerical computations by Chakraborty *et al* (1980).

2.4.1.1.3. Number of singletons (k_s).

An exact solution of (2.19) may be obtained for singletons or single copy alleles in the sample (k_s). By taking the binomial sampling equation for j=1, we get

$$E(k_s) = 2n \int_0^1 x(1-x)^{2n-1} \Phi(x)dx$$

$$= 2n\theta \int_0^1 (1-x)^{2n+\theta-2}dx$$

$$= 2n\theta/(2n+\theta-1) \qquad (2.40)$$

which yields a method of moment estimator, $\hat{\theta}_{k_s}$, as

$$\hat{\theta}_{k_s} = \frac{k_s(2n-1)}{(2n-k_s)} \qquad (2.41)$$

where k_s is the observed number of singletons/locus in the sample. For infinite sites model

$$\hat{\theta}_{1k_s} = k^*_s/2n \qquad (2.42)$$

where k^*_s is the proportion of sites segregated as singletons. The equation (2.42) is quite similar to (2.41) for large 2n.

The above solutions of the sampling equations have been given on the assumption of sampling from an infinite population with random mating, without replacement. This is not approached in actual situations. While the sampling procedures for finite populations will be taken up in the next section, it may be relevant to include the role of inbreeding in recovering the number of alleles in the population.

Templeton (1980) has given the appropriate expected number of neutral alleles under the infinite model, as

$$E(k|\alpha) = \int_0^1 [1-[1-x]^2+\alpha x(1-x)]^{2n}\Phi(x)dx$$

$$= E(k|\alpha=0) - \sum_{j=0}^{2n-1} \theta \frac{\alpha^{2n-j}}{2n-j} \frac{\Gamma(2n+1)\Gamma(2n+j+\theta)}{\Gamma(j+1)\Gamma(2n+\theta)} \qquad (2.43)$$

where $E(k|\alpha=0)$ is equation (2.11); $2n\Gamma$ and θ are already defined. It is readily seen from (2.43) that the mean number of alleles recovered in a sample decreases with increase in α, the inbreeding coefficient of the population. No simple expression for θ in terms of $E(k|\alpha)$, leading to an estimator based on k is, however, available.

In an earlier section it was pointed out that $\hat{\theta}_F$ is a biased estimator of θ. For $0.6 \leqslant \theta \leqslant 2.0$, Ewens and Gillespie (1974) show, by simulation, that the mean value of $\hat{\theta}_F$ is rather consistently about 40% or more in excess of θ. Although $\hat{\theta}_k$ is also a biased estimator (no unbiased estimator of θ exists; Ewens, 1979a), its bias decreases to zero asymptotically. For $2n=200$, the bias is negligible (Ewens, 1979a). Furthermore, $\hat{\theta}_k$ has very small mean square error typically 1/7th or 1/8th of $\hat{\theta}_F$. In the context of strict neutrality there appears no excuse for estimating θ through $\hat{F}$.

Under generalized neutrality, however, the above comparison does not hold universally for the large values of $\alpha'=2Ns$, where s is the selection coefficient. For large α', $\hat{\theta}_F$ has less bias than $\hat{\theta}_k$ under a number of conditions. It might seem paradoxical then to estimate θ from $\hat{\theta}_k$ when α' is low and from $\hat{\theta}_F$ when α' is high for the property of unbiasedness. In addition, the two estimates will then represent "total" θ and "neutral" θ.

In dealing with protein data, when strict neutrality is presumed but not either proved or disproved using a statistical test, the choice of $\hat{\theta}_F$ may fortiutously, provide a less biased estimate if the bias is removed by defining a new estimator $\hat{\theta}_F^*$, defined by

$$\hat{\theta}_F^* = 0.71\ \hat{\theta}_F \qquad (2.44)$$

This new estimator is designed to allow for the 40% bias in $\hat{\theta}_F$. Thus $\hat{\theta}_F^*$ will be an approximately unbiased estimator of 'total' θ i.e. when $\alpha'=0$ and an unbiased estimator of "neutral only" θ, whichever might apply. This blind estimation procedure is the only alternative, unless a strong test statistic are developed to discriminate between the various aspects of neutrality.

2.4.2 Branching process models.

Alternative approaches to estimate the mutation rate from protein data were suggested by Neel and Thompson (1978) and Rothman and Adams (1978) using branching process models. Fisher (1930) and Karlin and McGregor (1967) had earlier considered these models for estimating the total number of heterozygous loci in the genome and number of alleles represented by j copies in the population, respectively.

One of the advantages in working with branching process models is that the assumptions of fixed population size and equilibrium are not required to describe the transition of one allele from the jth allelic state to ith allelic state where an allelic state is defined by the number of copies by which an allele is represented in the population. However, for the estimation of mutation rate these questions are difficult to avoid.

2.4.2.1. Rothman and Adams' method.

Let K_s^t denote the number of different single copy alleles segregating at a locus in a population in tth generation. The presence of these alleles can be attributed to three different sources, provided there is no immigration/emigration and intragenic recombination is low. These sources are:

(1) New mutations introduced in the tth generation at a rate $2N^t\mu$ where N^t is the population size in the tth generation. It is assumed that an infinite series of alleles can be generated at this locus i.e. every new allele is a novel allele,

(2) The drift of higher frequency alleles in the t-1th generation to the singleton class in the tth generation,

(3) The retention of singletons in the t-1th generation as singletons itself in the tth generation.

The drift to and retention of singletons is given by the probability transition matrix P where individual elements of the matrix P_{ji} indicate the probability that an allele present in j copies in the t-1th generation is changed to i copies in the tth generation. Quantitatively, this is given by Rothman and Adams (1978) as

$$E(K_s) = 2N\mu + K \sum_{j=1} g(j) P_{ji} \qquad (2.49)$$

where $E(K_s)$ is the expected number of singletons in the population, g(i) is the relative proportion of alleles each represented by j copies in the populations, K is the expected number of alleles in the population and P_{ji} is the transition probability vector. This equation represents the balance, at equillibrium between the expected number of alleles entering the singleton class and those alleles which exit.

The method of Rothman and Adams of course, assumes that the mutational events are given under the infinite alleles model. The form of the equation (2.49) implicitly also assumes that mutation is introduced as a replication error during gametogenesis and is expressed phenotypically in the offspring. This being a unique event under the infinite alleles model, the possibility of a similar slip occuring again in the gametes of the parents is negligible.

An alternative model for the occurrence of mutation has been put forward by Vogel (1970; 1975). Under this model the mutation is introduced in the non-expressible form in the gamete cells of one of the parents. The probability of transmission of such mutations is governed by the usual demographic processes. The form of equation (2.45) under this model will be

$$E(K_s^t) = [2N^{t-1}\mu + K_s^{t-1}]\, P_{11} + K^{t-1}\sum_{j=2} g(j)P_{j1}$$

$$= 2N^{t-1}\mu P_{11} + K^{t-1}\sum_{j=1} g(j)P_{j1} \qquad (2.50)$$

Expanding over ℓ generations, and after rearranging we get at equilibrium:

$$E(K_s) = 2N\mu P_{11} + K\sum_{j=1} g(j)P_{j1} \qquad (2.51)$$

The model derived here, however, assumes that the mutations are introduced during the pre-pubertal period. Adjustments to the transmission probability P_{11} (associated with fresh mutations) will have to be made if the mutation is introduced in the gametes of the parents during the reproductive period.

Although the second model is not entirely acceptable (Vogel, 1975), the above equations have implications when the model is extended to expanding or contracting populations. While under the first model the mutation rate is measured in terms of the

size of the tth generation population size, under the second model the size of the previous generation is taken into consideration. A comparison of equations (2.45) and (2.51) reveals that, under the first model, the adjustment for the size of the previous generation is not admissible.

Rothman and Adams (1978) have given the equation which takes into consideration the growth rate per generation in the estimation of K_s. Accordingly,

$$K_s^t = 2N^{t-1}\mu + K^t \sum_{j=1} g(j) P_{j1} \qquad (2.52)$$

which is an extension of the approach taken by Lea and Coulson (1949). However, this equation is not extendable to any of the two models of mutation mentioned earlier.

Neel and Rothman (1978) rewrite the expression (2.49) as

$$2N\mu + K \sum_{j>1} g(j) P_{j1} = K(1)(1-P_{11}) \qquad (2.53)$$

which expresses the balance between the number of singletons lost and gained per generation. This equation has as unknowns, besides μ, the quantities P_{j1}, $g(j)$ and K.

The elements of transition probability matrix P are calculated as

$$P_{ji} = \sum_{h=1}^{\min(i,j)} \binom{j}{h}\binom{i-1}{i-h}\left(1-\frac{b}{1-c}\right)^{j-h} b^{h} c^{i-h} \qquad (2.54)$$

where b and c are the parameters of geometric distribution.

The population values of g(i), the expected relative frequencies of the alleles, are obtained as

$$\sum_{j=1} g(j)P_{j1} = g(i) \qquad (2.55)$$

for i>2. The relative frequency of g(1), however, is given as

$$g(1) = \sum g(j)P_{j1} + 2N\mu/K \qquad (2.56)$$

The estimation of the relative proportions g(j), however, needs a well documented demographic data on the population as also extensive computations. In the absence of such data, the rough estimates of g(j) can be obtained from the observed distribution of rare alleles by taking the number of copies over a set of protein loci for sufficient sample sizes.

Using hypergeometric sampling, Rothman and Adams

(1978) give the estimated number of alleles in the population, using the number k in the sample, as

$$\hat{K} = \frac{k}{\{1 - \sum_{j=1}^{2N} g(j) \frac{\binom{2N-j}{2n}}{\binom{2N}{2n}}\}} \qquad (2.57)$$

the binomial approximation of which is

$$\hat{K} = k/[1 - \sum_{j=1}^{2N} g(j)(1-f)^{j}] \qquad (2.58)$$

where $f = n/N$ is the sampling fraction. For $j \geq 30$, $g(j)(1-f)^{j}$ is negligible and the summation may be truncated.

The estimation procedure of Neel and Rothman (1978), however, is very difficult to utilise since there are too many unknowns. In the absence of well documented demographic data over a number of generations, calculation of the values of the elements of probability transitions matrix is difficult. Similarly the population values of g(j) are not known. Furthermore extrapolations of the values of k to obtain $\hat{K}$ is a very uncertain proposition since k is a random variable rather than an expected value.

2.4.2.2. Neel and Thompson's method.

Thompson and Neel (1978), using the tth generation distribution forms of the number of copies given by Keiding and Nielsen (1975) have given the parameters of cumulative distribution for the two-parameter-geometric form. Neel and Thompson (1978) utilize these results to give an estimator of mutation rate as

$$K_A(A \geq j) = \mu N \sum_{t}^{T} \{(1-\frac{1}{W_t})^j \frac{1}{W_t}\} \qquad (2.59)$$

where K_A represents the number of alleles with more than or equal to j copies in the population/locus, and W_t is the tth generation mean value of replicates conditional on non-zero. However, the summation on the right hand side is unbounded which, for the private variants, may be truncated to include only the time since tribal differentation. The approach is quite useful for utilizing information on private polymorphisms.

2.4.3. Equilibrium methods.

The diffusion approximations approach outlined above helps in arriving at some of the results in simple approximate forms. These approximations are, however, based on a number of assumptions which may be considered unrealistic for natural populations.

Included in this section is the equilibrium approach of Ewens (1964) and Kimura and Ohta (1969),

the forms of which are estimated by using diffusion approximations. The details of this method are already given in Chapter 1.

Chapter 3

THE FREQUENCY OF PRIVATE ELECTROPHORETIC VARIANTS AND INDIRECT ESTIMATES OF MUTATION RATE

3.1. Introduction.

The present chapter makes use of some of the methods outlined in Chapter 2 for indirectly estimating the rate of mutation from electrophoretic data for various world populations. Some of the results included in Section 3.2 have appeared already in a series of papers: Bhatia *et al.* (1979) on Australian Aborigines; Bhatia *et al.* (1981) on Papua New Guineans and Bhatia (1981b) on some Scheduled Tribe populations from India. These results are presented here with some modifications. In addition, estimates have been generated on some Khoisan and Negro populations of southern Africa and included in Section 3.2.4.

Since the above mentioned papers were written another estimation procedure has been suggested by Chakraborty (1981). In addition, some other aspects of the allelic data (e.g. singletons) obtained through sampling, can be used to generate estimates of mutation rate. To update the results, further estimates of mutation rate on Australian Aborigines, Papua New Guineans and tribal populations from India have been made and are given in Section 3.2.5.

For each of these populations a brief resume of the population is given, followed by the listings of the laboratory data and the parameters required in estimating the rate of mutation by the various methods used. The estimates of mutation rate in each population are then discussed individually.

3.2. Estimates of Mutation Rate.

3.2.1. Mutation rate in Australian Aborigines.

3.2.1.1. The study population.

At the time of first European contact, the Aborigines were spread across the Australian continent, having exploited, with few exceptions, all the available ecological situations. Their presence in the continent is dated back to at least 40,000 years, though the occupation of the more arid areas in the center probably took place no more than 10,000 years ago (Kirk, 1981). At the time of European contact the population of Aborigines has been estimated at about 250,000 (Radcliffe-Brown, 1930), and the population was divided into several hundred tribal and local groups varying in size from 100 to several thousand persons (Tindale, 1974).

During the last 200 years the Aboriginal population of Australia fell dramatically, reaching its lowest reported level in the census of 1921. This population decrease was not uniform; in some areas such as Tasmania, the eclipse was total while, in many others across the southern portion of the

continent, there are few, if any, persons of full Aboriginal descent remaining. In areas more remote from European settlement the decline in numbers was less, but even here the total may have been reduced to 50% before the increase in population characterizing the present situation commenced. The present analysis is based on samples from this area of minimum disturbance shown in figure 3.1.

There are no accurate records of the age structure in traditional Aboriginal populations (Smith, 1980). Available data refer to populations already exposed to varying degrees of European contact. At present, the age structure for persons of full Aboriginal descent shows a heavy-based pyramid with only 41.6% in the 15-44 years age group (Commonwealth of Australia, 1975). In the traditional situation, each population may have varied in demographic parameters influenced by natural disasters such as prolonged drought or cyclones. Such factors may have led to drastic reductions in number followed by subsequent population expansion or by replacement through migration from neighbouring groups. Over a longer time period, however, we assume that the population of the continent was in equilibrium, and that the average net increase was zero.

Since the precise boundary of the total Aboriginal population in our surveys is difficult

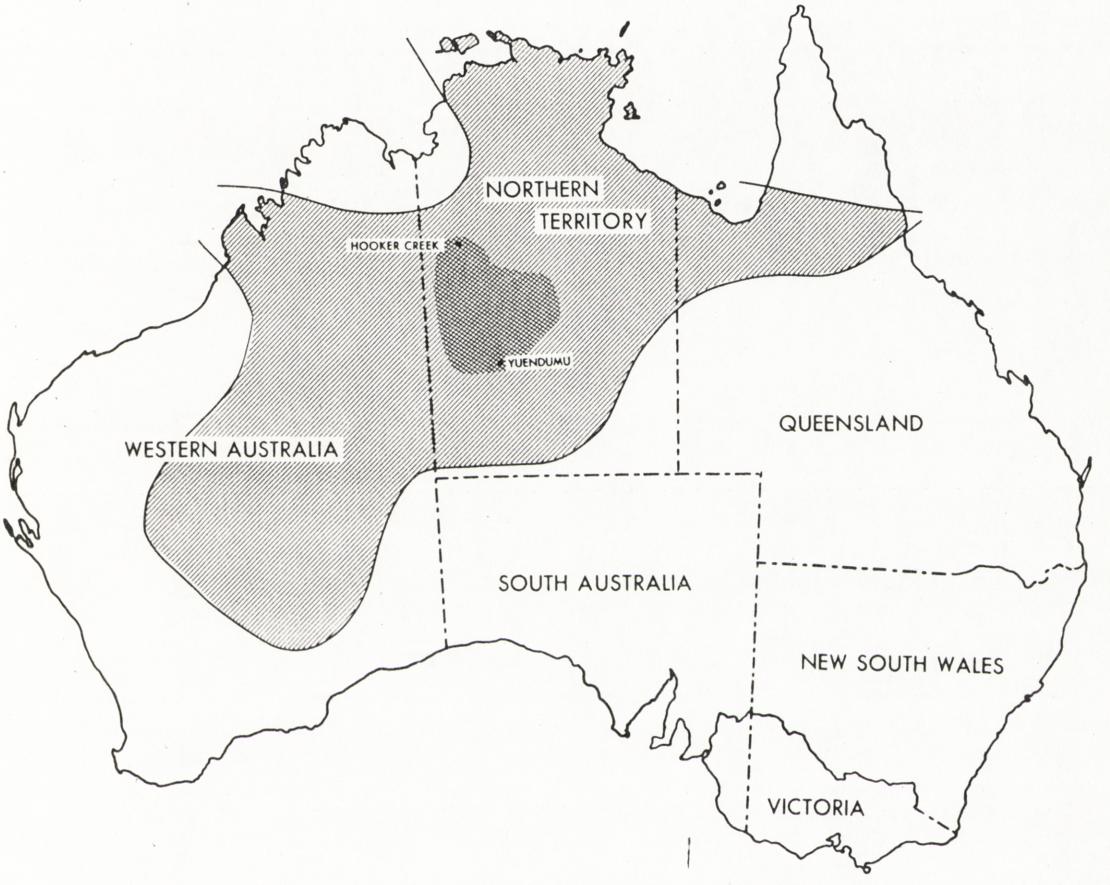

Fig. 3.1 Map of Australia showing areas sampled (diagonal hatching) and tribal territory of the Waljbiri (cross hatched).

to define, I have provided data also for one specific tribal group, defined by the spoken language Waljbiri, one of the largest linguistic groups in the Northern Territory. The Waljbiri territory (see fig.3.1) covers 35,000-40,000 square miles of arid and semi-arid country, and the population density averages one person per 25-27 square miles (Meggitt, 1962).

Meggitt's detailed study of the Waljbiri revealed that the dialectical Waljbiri tribe is further divided into four subgroups, namely, Yalpari (Lander), Waneiga, Walmalla and Ngalia. Marriages between the subgroups are frequent, according to Meggitt. Tindale (1953), however, found only 1.3% marriages between Ngalia and Walmalla, and no Yalpari-Waneiga marriages were recorded. Birdsell (1970) claims that before 1935, the Ngalia subgroup was quite distinct from the other Waljbiri. Intertribal marriages involving Ngalia, however, were significantly higher at 6%-7%.

The Waljbiri in this series were sampled mainly at two localities, Yuendumu and Hooker Creek. The Yuendumu Waljbiri predominantly belong to the Ngalia subgroup, though some reside also at Hooker Creek. Although I have pooled the results for all Waljbiri, the data indicate a clear-cut heterogeneity between the populations

at these two localities.

3.2.1.2. The laboratory data.

This analysis is confined to data for red cell enzyme proteins and haemoglobin, representing products of genes at 25 loci. The basic data have been tabulated recently by Blake (1979) and are summarized in table 3.1. A total of 16 detected variants restricted to Australian Aborigines are listed in table 3.2, together with the number of copies observed and their gene frequencies. Three of the variant alleles (PGM_2^3, CA_1^9, and CA_2^4) have achieved frequencies above 1% and can be classified as polymorphic. Two others (PGD^{Elcho} and $PEP\ B^6$) have allele frequencies approaching 1%, and the remainder are more restricted, the number of copies ranging from one to 14. Table 3.2 also shows separately the number of rare variants detected in the Waljbiri tribe. Only five of the 16 rare variants among Aborigines were detected among the Waljbiri, four of these being polymorphic in this tribe, while the other ($PEP\ B^6$) has an allele frequency of 0.74%. Three of the polymorphic alleles among the Waljbiri are polymorphic in Aborigines in general. In the case of the other peptidase variant allele ($PEP\ B^7$), 13 of the 14 copies occurred among Waljbiri, the other example being found in Luridja, a group known to intermarry with the Waljbiri.

TABLE 3.1 GENETIC MARKERS IN AUSTRALIAN ABORIGINES

(Based on Blake, 1979)

LOCUS NO.	ENZYME SYSTEM	ABBREVIATION	SAMPLE SIZE
1	6-Phosphogluconate dehydrogenase	6PGD	4035
2	Acid phosphatase-1	ACP_1	4016
3	Phosphoglucomutase-1	PGM_1	3919
4	Phosphoglucomutase-2	PGM_2	3790
5	Peptidase A	PEPA	3034
6	Peptidase B	PEPB	3189
7	Carbonic anhydrase-1	CA_1	3751
8	Carbonic anhydrase-2	CA_2	3751
9	Glyoxylase	GLO	1290
10	Adenosine deaminase	ADA	1437
11	Esterase D	EsD	1556
12	Glutamic pyruvic transaminase	GPT	1391
13	Hemoglobin-α	Hbα	2692
14	Hemoglobin-β	Hbβ	2692
15	Diaphorase	DIA	1861
16	Glucose-6-phosphate dehydrogenase	G-6-PD	1014
17	Malate dehydrogenase-2	MDH_2	2964
18	Superoxide dismutase	SOD	1795
19	Lactate dehydrogenase-A	LDH_A	4180
20	Lactate dehydrogenase-B	LDH_B	4180
21	Isocitrate dehydrogenase	ICD_s	1226
22	Phosphohexose isomerase	PHI	1569
23	Adenylate kinase-1	AK_1	3535
24	Phosphoglycerate kinase	PGK	1569
25	Glutamic oxaloacetic acid transaminase	GOT	748

TABLE 3.2 THE NUMBER AND FREQUENCIES OF PRIVATE VARIANTS IN AUSTRALIAN ABORIGINES

SR. NO.	ENZYME	VARIANT	TOTAL POPULATION		WALJBIRI	
			Number of Copies	% Gene Frequency	Number of Copies	% Gene Frequency
1	6-PGD	PGD^{Elcho}	65	0.81	0	0.00
2	PEPA	$PEP\ A^{3}$	1	0.02	0	0.00
3	PEPB	$PEP\ B^{6}$	53	0.83	6	0.74
4	PEPB	$PEP\ B^{7}$	14	0.21	13	1.43
5	PGM_1	PGM_1^{6}	4	0.05	0	0.00
6	PGM_1	PGM_1^{7}	1	0.01	0	0.00
7	PGM_2	PGM_2^{3}	103	1.36	46	5.68
8	PGM_2	PGM_2^{11}	6	0.08	0	0.00
9	ACP_1	ACP_1^{F}	1	0.01	0	0.00
10	CA_1	CA_1^{9}	192	2.53	36	4.44
11	CA_1	CA_1^{10}	8	0.11	0	0.00
12	CA_2	CA_2^{4}	166	2.21	14	1.73
13	LDH_B	LDH_B^{SLOW}	1	0.01	0	0.00
14	LDH_A	LDH_A^{SLOW}	2	0.02	0	0.00
15	G-6-PD	Gd_B^{FAST}	1	0.04	0	0.00
16	PHI	PHI^{4}	1	0.03	0	0.00

3.2.1.3. Estimation of k_t, k_r and K.

For the total Aboriginal population, 25 enzyme loci have been studied. Of these, 13 showed no polymorphism by the standard definition where the least common allele frequency did not exceed 1%, but private variants were detected at four of these loci. Twelve private variants were distributed among eight of the 12 polymorphic loci. The mean sample size (table 3.3) for the polymorphic loci without private variants (1,419) is significantly lower than those with private variants (3,686). In my data, therefore, the probability of detecting private variants among known polymorphic loci increases with sample size.

The data in table 3.3 clearly show that the value of $\hat{k}$ varies also with the type of loci (polymorphic or monomorphic). The probability of detecting private variants increases with sample size, which will also influence the value of $\hat{k}$.

In Nei's method (1977), k is calculated only from those private variants which are not polymorphic. The differences for all loci between $\hat{k}_t$ and $\hat{k}_r$ is 0.12 for the total Aboriginal population. For the Waljbiri population, the only private variant (*PEP* B^7) is polymorphic. Since it is an exclusive tribal marker for the Waljbiri, I have used it in the calculation of $\hat{k}_r$, giving a value of 0.04.

TABLE 3.3 NUMBER OF LOCI WITH AND WITHOUT PRIVATE VARIANTS AND VALUES OF k IN THE TOTAL ABORIGINAL POPULATION

	POLYMORPHIC LOCI			MONOMORPHIC LOCI			TOTAL LOCI		
	With Private Variants	Without Private Variants	Total	With Private Variants	Without Private Variants	Total	With Private Variants	Without Private Variants	Total
Number of loci	8	4	12	4	9	13	12	13	25
Number of private variants	12	0	12	4	0	4	16	0	16
$\hat{k}_t$ per locus	1.50	0.00	1.00	1.00	0.00	0.31	1.33	0.00	0.64
$\hat{k}_r$ per locus	1.12	0.00	0.75	1.00	0.00	0.31	1.08	0.00	0.52
$\hat{K}$ per locus	2.10	0.00	1.40	1.55	0.00	0.48	1.80	0.00	1.07
Mean sample size (per locus)	3686	1419	2930	2736	2120	2309	3369	1904	2607
Mean sample size (per variant)	2457	-	2930	2736	-	7506	2527	-	4074

The estimates of $\hat{K}$ obtained by using the binomial approximation sampling method of Rothman and Adams (1978) are 1.40, 0.48 and 1.07 variants per locus for polymorphic, monomorphic and total loci respectively. The values of $\tilde{g}(i)$ used in these extrapolation are based on the observed values. For Waljbiri tribal group we use the $\tilde{g}(i)$ values estimated for the total sample which gives an estimate of $\hat{K}$ as 0.056.

3.2.1.4. The estimation of actual (apparent) population size, N.

As explained earlier, with the data available, it is not possible to give a precise estimate of the actual population size because of changing reproductive patterns among Aborigines. Here I use the population in the 15-44 years age group in the 1961 census year, adjusted for the proportion of the total population in the surveyed area.

The population of full descent Aborigines in Australia in 1961 was 36,137 (18,899 males; 17,238 females), of which 41.6% were in the age cohort 15-44 years (Commonwealth of Australia, 1975), and the area surveyed contains approximately 60% of the total full descent population. This gives a value of N=9,160.

It can be argued that this does not represent the effective population size of Australian

Aborigines during most of their stay on the continent. However, indirect evidence suggests the difference in age structure in traditionally oriented societies is not likely to be very different from the value used here. For example. Tindale (1974) has recorded approximate age composition for three nomadic bands encountered in the central desert areas. The mean value for the adult composition of these bands is 28.0%. Since this covers the age range 20-40 years, the composition of the 15-44 cohort will not be very different from the 41.6% derived from the 1961 census. In the case of the Waljbiri I have age estimates for the Yuendumu population (Middleton and Francis, 1976). This gives 43.2% for the 15-44 age cohort. From the total Waljbiri population estimate given by Milliken (1976), N for Waljbiri becomes 1,173.

Another difficulty is that Aboriginal populations have been subject to a series of bottleneck effects due to the operation of various factors. This could cause the loss of a number of private variants which, in turn, will affect the calculation of $\hat{k}_t$, $\hat{k}_r$ and $\hat{K}$. The loss of these private variants, however, will be proportional to the decline in population size. On the other hand, private variants which survive the population crash will increase in number during the subsequent population expansion.

3.2.1.5. The estimation of $\bar{t}_o$.

Kimura and Ohta (1969) showed that the mean survival time for a neutral mutation in generations in terms of variance effective population size ($N_e v$) and actual population size N is given by

$$\bar{t}_o = 2 \frac{N_e v}{N} \log_e(2N)$$

in a stationary non-subdivided population with no reproductive death and the progeny size following a Poisson distribution. Kimura and Maruyama (1971), however, argue that if the population is subdivided into loose random mating units between which migration occurs, it may be treated approximately as a single random mating unit, disregarding the substructure of the populations.

Applying Kimura and Ohta's formula to the Aboriginal populations, and using the estimates of N given above, I obtain values of $\bar{t}_o$ ~12.42 generations for the Waljbiri. Neel and Rothman (1978), however, consider the values of $\bar{t}_o$ calculated by this method as overestimates. The mean survival time can be simulated for each population, and Li and Neel (1974) and Li *et al.* (1978) obtained values between 2.3 to 2.8 generations. However, after making concessions for the various factors influencing the population, Neel and Rothman (1978) give a mean value of 5.7 generations. I shall use this value here.

3.2.1.6. Results.

Mutation rates estimated by each of the three methods listed above, both for the total Aboriginal population surveyed and for the Waljbiri tribal group are given in table 3.4. The rates vary within a range from 3.58×10^{-6} to 12.72×10^{-6}, with a mean value of 8.85×10^{-6}/locus per generation. In obtaining the values of μ based on Kimura and Ohta (1969), the mean number of variants per locus were obtained for the actual population size.

Mutation rates estimated from private variants at polymorphic loci are 2-3 times higher than those estimated from the monomorphic loci. This may be a function of the smaller sample sizes for the private variants at monomorphic loci in our sample, which will have reduced the probability of detecting private variants. Eanes and Koehn (1978) recently have also drawn attention to the relationship between sample size and detection of rare electrophoretic variants.

The values of $\mu = 4.07 \times 10^{-6}$ to 5.23×10^{-6}/ locus per generation for the Waljbiri are lower than the values obtained for the total Aboriginal sample. These lower values are due to the fact that while five private variants were detected in the Waljbiri, only one is included for calculating, $\hat{k}_t$, $\hat{k}_r$ and $\hat{K}$. The other four are more widely

TABLE 3.4

MUTATION RATES ($x10^6$) IN AUSTRALIAN ABORIGINES

ESTIMATED BY VARIOUS METHODS

	Kimura and Ohta's Method	Nei's Method	Rothman and Adams' Method
	$\hat{\mu}_{K-O}$	$\hat{\mu}_{NEI}$	$\hat{\mu}_{R-A}$
Polymorphic loci Total Aboriginal sample	13.40	5.03	16.63
Monomorphic loci Total Aboriginal sample	4.60	2.20	5.71
All loci[1] Total Aboriginal sample	10.25	3.58	12.72
Waljbiri sample	4.22	4.07	5.23

[1] Mean = 8.85×10^{-6}

distributed in the Aboriginal population, and it is not possible to assign the original mutants to the Waljbiri.

Neel and Rothman (1978) estimated mean mutation rates based on values for 12 Amerindian tribes in South America by each of the same three methods. The unweighted mean for the 12 tribes averaged for the three methods is 16×10^{-6}/locus per generation. The mean value for the Amerindians is almost twice our value for the total Aboriginal sample. However, Neel and Rothman's value of 16×10^{-6} is based on unweighted means for the tribal samples. If it is recalculated using weights based on the effective population sizes, the weighted mean value becomes 7.2×10^{-6}/locus per generation. It is interesting to note that recently Neel and Thompson (1978), using a method based on simulation, arrived at a mean mutation rate of 7.0×10^{-6}/locus per generation. These values are very similar to our own based on the total Aboriginal sample. The value for the Waljbiri, of course, is only half that for the total Aboriginal sample. Neel and Rothman found a range of values of $0\text{-}51 \times 10^{-6}$/locus per generation for their 12 Amerindian tribes. The Waljbiri, therefore, fall within this range and we assume that, if data were available for a similar number of tribal populations in Australia,

the range of values may also be similar to those for the Amerindians.

Although the indirect estimation of mutation rates using data on private electrophoretic variants has many problems, ranging from the technical factors influencing the recognition of rare variants, through sampling design to the estimation of $\hat{k}_t$, $\hat{k}_r$, $\hat{K}$ and N, it is of great interest that data collected in two different laboratories from studies of different populations on two continents have yielded estimates of μ which are so similar.

3.2.2. Mutation rates in Papua New Guinean populations.

3.2.2.1. The study population.

Papua New Guinea comprises the portion of the island of New Guinea east of longitude 141°E together with several geographically related islands, including New Britain, the Admiralty Islands and Bougainville. The census size of Papua New Guinea is approximately three million, or about 67 per cent of the estimated total Melanesian population, and its population is one of the linguistically most complex and socially fragmented areas of the world. It is estimated that there are about 700 speech communities in Papua New Guinea divided among two major linguistic

phyla, Papuan and Austronesian (Wurm, 1975a). A survey of the patterns of social structure is given in the Encyclopaedia of Papua New Guinea (Lepervenche, 1972).

The distribution of population densities in Papua New Guinea is highly uneven. The highland region is one of the most densely populated areas in Papua New Guinea (14 persons/km^2 against 4.7 persons/km^2 in the country as a whole) and in some regions the density ranges from 100-150 persons/km^2 (Brown and Podolefsky, 1976). Approximately 40% of the total population lives in the Highlands and another 17% on offshore islands. Watson (1965a,b) attributes this high level of density in the highlands to the introduction and cultivation of sweet potato about three centuries ago, which resulted in explosive population growth. Some researchers do not agree with this explosion theory of population growth (Brookfield and White, 1968), yet the projected population of Papua New Guinea in the 17th century could not have been more than 225,000 (van de Kaa , 1971-72).

This population size of 225,000 should be the equilibrium population of Papua New Guinea at the hunter-gatherer level before the introduction of agriculture which, over a span of 12 generations, increased in population size about

tenfold. The age and sex composition of the population could, however, have been much different in the past from what it is now. The sex ratio in Papua New Guinea was 1,088 males per 1,000 females in 1966 (van de Kaa, 1971-72) with about 44.6% in the age group 15-44 years and the incidence of polygyny was about 10.0%. But the sex ratio in the 1971 census was less than unity (Administration of Papua New Guinea, 1972) with about 49% in the age group 15-44 years. These wide fluctuations in demographic features may, however, be associated with the recent population increase.

The present analysis is based on samples collected from populations belonging to 47 languages, also called speech communities, on the mainland of Papua New Guinea, together with two speech communities from Karkar Island and five from Siassi Islands, both off the northern shore of New Guinea and one speech community, Titan, from the Great Admiralty Island, also called Manus. The populations of these offshore islands, namely Karkar, Manus and Siassi, although having evolved in a similar ecological setting, have been exposed to different types of population pressures. These societies, like the coastal regions of mainland Papua New Guinea, have been at the crossroads of migrations, in and around the Pacific, and may well have had their genetic

composition considerably altered through repeated contact with outsiders. The populations of these three islands have been analyzed separately also for respective mutation rates.

The 55 speech communities included in the present study are listed in table 3.5, along with their estimated population size. A map depicting the position of various units is given in Figure 3.2.

3.2.2.2. The laboratory data.

The data analyzed here include only the published and unpublished genetic surveys conducted by the members of the Department of Human Biology in collaboration with other workers. Material in all cases was shipped by air to Canberra and laboratory testing carried out in our laboratories using standard procedures outlined in Blake *et al.* (1973). The variants of a few other systems were tested using the techniques as follows: GPT (Chen *et al.* , 1972), EsD (Hopkinson *et al.* 1973), CA_1 and CA_2 (Hopkinson *et al.*, 1974), GLO (Kompf *et al.*, 1975), PGM_1 and PGM_2 (Blake and Omoto, 1975). The list of enzymes studied, along with their sample size and subunit molecular weights, are given in table 3.6.

Seven of the 21 loci included in the study are polymorphic and six of the 21 are invariant. Out of the 53 alleles segregating 24 alleles are rare as a whole. Two other alleles, namely PGM^3_1

TABLE 3.5 SPEECH COMMUNITIES SAMPLED IN PAPUA NEW GUINEA

ADMINISTRATIVE DISTRICT	LANGUAGE FAMILY*		LANGUAGE	ESTIMATED SIZE OF SPEECH COMMUNITY	REFERENCE
Central	Group II	A	Motu	13,000	Taylor, 1970.
	Mailuan	NA	Mailu	4,662	Dutton, 1971
	Goilalan	NA	Fuyuge	9,615	Steinkraus and Pence, 1964
Milne Bay	Dagan	NA	Daga	5,326	Dutton, 1971
Eastern Highlands	Eastern	NA	Gadsup	9,100	Gajdusek and Alpers, 1972
			Asaro	12,000	-ditto-
			Benabena	12,300	-ditto-
			Fore	15,100	-ditto-
			Agarabi	11,000	Wurm, 1975b
Chimbu	Central	NA	Kuman (Chimbu)	66,000	-ditto-
Western Highlands	W. Central	NA	Enga	150,000	-ditto-
Morobe	Angan	NA	Menyama Anga'	12,400	-ditto-
	Wantoat	NA	Wantoat	5,000	Classen and McElhanon, 1970
			Irumu	1,800	-ditto-
			Waritsian**	500	Department of Chief Minister and Development Admin. 1970-74
Morobe (Siassi Is.)	Siassi	A	Barim	325	-ditto-
			Tuam Mutu	3,330	-ditto-
			Lukep	627	-ditto-
			Mangap	2,635	-ditto-
	Kovai	NA	Kovai	2,108	-ditto-
Manus (Manus Is.)	Manus	A	Titan	2,550	Healey, 1975
Madang (Karkar Is.)	Siassi	A	Takia	10,962	Z'Graggen, 1975a
	Kowan	NA	Waskia	530	Z'Graggen, 1975b
Madang	Yupna	NA	Eandabong	9,106	McElhanon, 1975
			Nokopo	1,690	-ditto-
			Kewieng	940	-ditto-
Madang	Kokon	NA	Bemal	642	Z'Graggen, 1975b
			Munit	345	-ditto-
	Amaimon	NA	Amaimon	701	-ditto-
	Kalam	NA	Gants	1,900	Wurm, 1975b
	Gum	NA	Sihan	314	Z'Graggen, 1975b

TABLE 3.5 Cont'd

ADMINISTRATIVE DISTRICT	LANGUAGE FAMILY *		LANGUAGE	ESTIMATED SIZE OF SPEECH COMMUNITY	REFERENCE
Madang cont'd	Belan	A	Ham	1,495	Z'Graggen, 1975a
	Meseman	A	Manam	5,950	-ditto-
			Sepa	268	-ditto-
	Kaukombaran	NA	Pay	769	Z'Graggen, 1975b
			Pila	669	-ditto-
			Saki	2,403	-ditto-
			Tani	2,494	-ditto-
	Miseigan	NA	Mikarew	5,872	Laycock and Z'Graggen, 1975
			Sepen	428	-ditto-
			Giri	1,819	-ditto-
	Ataitan	NA	Tangu	2,684	-ditto-
	Monumbo	NA	Monumbo	450	Laycock, 1975
			Lilau	410	-ditto-
	Wadaginam	NA	Wadaginam	546	Z'Graggen, 1975b
	Pomoikan	NA	Moresada	197	-ditto-
E. Sepik	Ndu	NA	Iatmul	9,842	Laycock, 1975
	Sepik Hill	NA	Alamblak	1,107	-ditto-
			Kapriman	1,439	-ditto-
			Sumariup	65	-ditto-
	Pondo	NA	Karawari	1,300	-ditto-
Gulf	Purari	NA	Purari	6,500	Wurm, 1975b
W. Sepik	Ok	NA	Tifal	2,500	Voorhoove, 1975
Western	Ok	NA	Kauwal	500	-ditto-
Southern Highlands	Bosavi	NA	Onabasulu	300	-ditto-
Total				416,215	

* A, Austronesian; NA, Non-Austronesian or Papuan.

** A genetic isolate

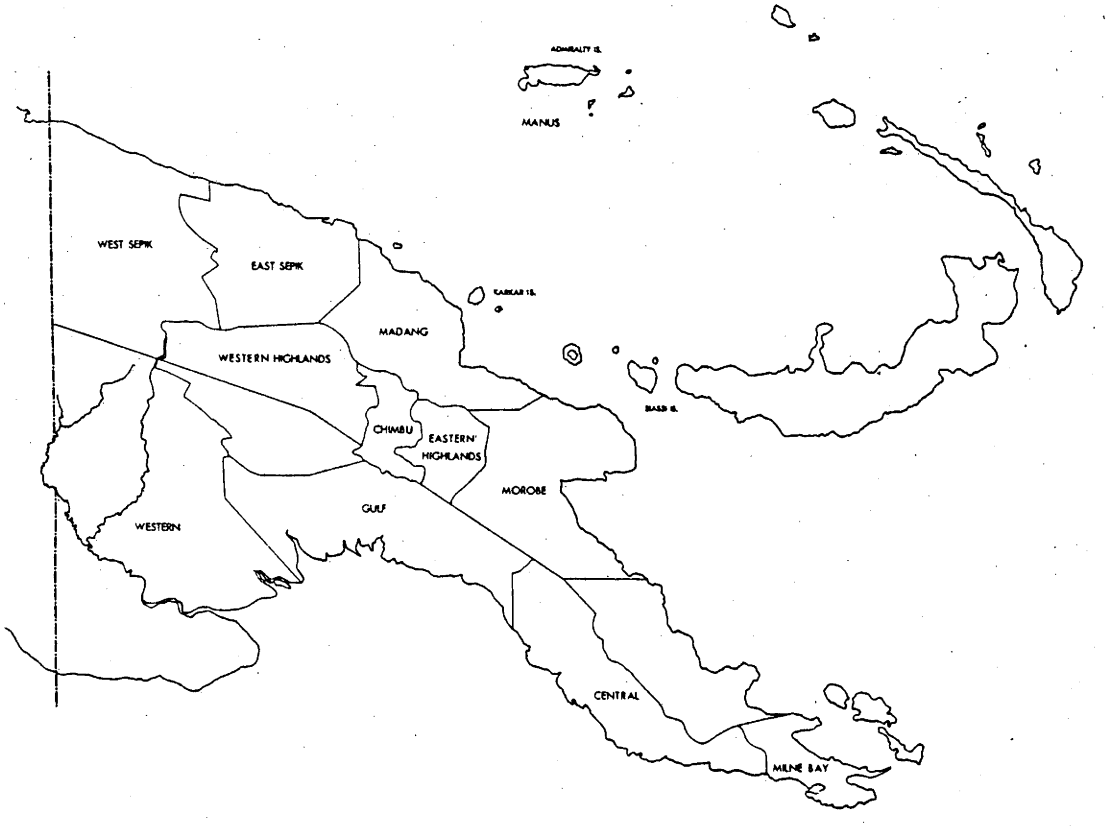

Fig. 3.2 Map of Papua New Guinea showing boundaries of administrative areas and island populations included in the present investigation.

TABLE 3.6 GENETIC MARKERS IN PAPUA NEW GUINEANS

LOCUS NO.	ENZYME SYSTEM	ABBREVIATION	SUBUNIT SIZE (IN DALTONS)	SAMPLE SIZE			
				Total Population	Karkar Island	Manus Island	Siassi Islands
	Multimers:						
1	Hemoglobin-α	Hbα	15,000	6,874	1,086	80	287
2	Hemoglobin-β	Hbβ	16,000	6,874	1,086	80	287
3	Superoxide dismutase	SOD_A	16,000	7,322	1,091	183	287
4	Glyoxalase I	GLO	24,000	2,192	-	-	283
5	Esterase D	ESD	28,000	5,453	870	-	286
6	Malate dehydrogenase-2	MDH_2	35,000	7,856	1,091	183	286
7	Lactate dehydrogenase A	LDH_A	35,000	7,858	1,091	183	286
8	Lactate dehydrogenase B	LDH_B	35,000	7,858	1,091	183	286
9	Glutamic-oxaloacetic transaminase	GOT	46,000	4,441	433	-	-
10	Isocitrate dehydrogenase	ICD_S	48,000	4,908	433	-	287
11	Glutamic-pyruvic transaminase	GPT	50,000	5,205	837	-	93
12	6-Phosphogluconate dehydrogenase	6PGD	53,000	7,809	1,091	183	284
13	Phosphohexose isomerase	PHI	62,000	6,934	617	183	221
	Monomers:						
14	Acid phosphatase-1	ACP_1	15,000	7,421	1,090	183	287
15	Adenylate kinase-1	AK_1	22,000	5,551	617	183	-
16	Carbonic anhydrase-1	CA_1	29,000	1,857	-	-	287
17	Carbonic anhydrase-2	CA_2	29,000	1,370	-	-	-
18	Phosphoglycerate kinase	PGK	50,000	7,818	1,090	183	260
19	Phosphoglucomutase 1	PGM_1	51,000	7,775	1,087	183	283
20	Peptidase B	PEPB	55,000	5,108	617	183	-
21	Phosphoglucomutase 2	PGM_2	61,000	7,787	1,086	183	283
TOTAL				126,271	16,404	2,356	4,573

and PGM_2^9, although polymorphic, are here considered to be of New Guinea origin. This raises the number of variants to 26. The names of these variants, along with their frequencies are given in table 3.7.

Four of these 26 variant alleles, namely *Hb* $J^{Tongariki}$, GOT^3, GPT^3 and GPT^6 cannot be assigned with certainty to Papua New Guinea because of their presence in appreciable numbers in other western Pacific populations which, as is true for the Japanese during World War II, have had contact with Papua New Guinea in the past. The introduction of these variants to Papua New Guinea, through admixture, therefore, cannot be excluded. Thus we have 22 alleles at 21 loci which can be regarded as indigenous to Papua New Guinea.

In Karkar Island we detected 10 of the 26 allelic variants listed in table 3.7 over a set of 18 enzyme loci. Only two of these, namely LDH_B^{KK2} and LDH_B^{KK3} are unique to Karkar and are represented by single copies only. Seven of the other eight variants are mainland markers also; the only exception is *Hb* $J^{Tongariki}$, polymorphic in Karkar but rare on the mainland. This last allele also has wide distribution in other parts of Melanesia.

No private variant was detected at 14 red cell enzyme loci tested in Manus. The absence of

TABLE 3.7 NO. AND FREQUENCIES OF PRIVATE VARIANTS IN PAPUA NEW GUINEA

Serial No.	Enzyme	Private Variant	TOTAL POPULATION		KARKAR ISLAND		MANUS ISLAND		SIASSI ISLAND	
			No. Copies	% Gene Frequency	No. Copies	% Gene Frequency	No. Copies	% Gene Frequency	No. Copies	% Gene Frequency
1	Hbα*	*Hb J*Tongariki	46	0.34	38	1.75	0	0.00	7	1.22
2	ESD	ESD^{3}	1	0.01	0	0.00	-	-	0	0.00
3	MDH_2	MDH_2^{3}	106	0.68	16	0.73	0	0.00	0	0.00
4	MDH_2	MDH_2^{6}	1	0.01	0	0.00	0	0.00	0	0.00
5	LDH_A	LDH_AWantoat	1	0.01	0	0.00	0	0.00	0	0.00
6	LDH_A	LDH_A^{WH}	9	0.06	0	0.00	0	0.00	0	0.00
7	LDH_B	LDH_B^{KK2}	1	0.01	1	0.05	0	0.00	0	0.00
8	LDH_B	LDH_B^{KK3}	1	0.01	1	0.05	0	0.00	0	0.00
9	GOT*	GOT^{3}	1	0.01	0	0.00	-	-	-	-
10	GPT*	GPT^{3}	8	0.08	0	0.00	-	-	0	0.00
11	GPT*	GPT^{6}	10	0.10	3	0.18	-	-	0	0.00
12	6PGD	$PGD^{Wantoat}$	38	0.24	3	0.14	0	0.00	0	0.00
13	6PGD	$PGD^{Hackney}$	1	0.01	0	0.00	0	0.00	0	0.00
14	PHI	PHI^{2}	1	0.01	0	0.00	0	0.00	1	0.23
15	PHI	PHI^{3}	1	0.01	0	0.00	0	0.00	0	0.00
16	PHI	PHI^{5}	4	0.03	0	0.00	0	0.00	0	0.00
17	PHI	PHI^{9}	2	0.02	0	0.00	0	0.00	0	0.00
18	PHI	PHI^{11}	1	0.01	0	0.00	0	0.00	0	0.00
19	PGK	PGK^{2}	18	0.12	0	0.00	0	0.00	11	2.12
20	PGK	PGK^{4}	62	0.40	5	0.23	0	0.00	3	0.58
21	PGM_1	PGM_1^{3}	254	1.63	17	0.78	0	0.00	1	0.18
22	PGM_1	PGM_1^{6}	3	0.02	0	0.00	0	0.00	0	0.00
23	PEPB	$PEPB^{2}$	1	0.01	0	0.00	0	0.00	0	0.00
24	PGM_2	PGM_2^{3}	1	0.01	0	0.00	0	0.00	1	0.18
25	PGM_2	PGM_2^{9}	903	5.80	166	7.64	0	0.00	10	1.77
26	PGM_2	PGM_2^{10}	112	0.72	4	0.18	0	0.00	3	0.53

Southwest Pacific genetic markers like PGM_1^3, PGM_1^7, PGM_2^9, PGM_2^{10}, PGK^2, PGK^4 and *Hb* $J^{Tongariki}$, makes this population unique in the western Pacific area.

In a set of 17 loci, two unique variants were found in Siassi Islands. These are PHI^2 and PGM_2^3, with a single copy each. The population of Siassi Islands, however, has a fair proportion of markers distributed in the western Pacific region. Except for a very low frequency of PGM_1^3, four other variants namely, PGK^2, PGK^4, PGK_2^{10} and *Hb* $J^{Tongariki}$ are in polymorphic proportions in these islands.

3.2.2.3. Estimation of k_t, k_r and K.

Three different estimates for the mean number of variants/locus were calculated. For the 21 loci included in the present study, 22 unique (20 rare and 2 polymorphic) variants were detected which give values of $\hat{k}_t$, $\hat{k}_r$ and $\hat{K}$ as 1.05, 0.95 and 2.10 respectively. The two parameters of geometric distribution involved in the estimation of K, namely b and c ($\hat{b}$=0.5567; $\hat{c}$=0.3865) (Rothman and Adams, 1978) have been estimated from the data for Kiunga in the Western Province given by Serjeantson (1975).

These estimates, however, are underestimates if the role of various factors affecting the incidence and detection of rare variants is considered (table 3.8). For example, the minimum sample size for loci with variants is 4,441. Three of

TABLE 3.8 MEAN SAMPLE SIZES, SUBUNIT MOLECULAR WEIGHTS AND HETEROZYGOSITIES FOR PROTEIN LOCI WITH AND WITHOUT PRIVATE VARIANTS

TYPE OF LOCI	TYPE OF ENZYME	NUMBER OF LOCI	TOTAL NUMBER OF ALLELES	TOTAL NUMBER OF PRIVATE ALLELES	MEAN SAMPLE SIZE	MEAN SUBUNIT MOLECULAR WEIGHT	MEAN HETEROZY-GOSITY $(1-\Sigma\chi_1^2)$
With	Multimers	6	22	14	7,295	41,167	0.0538
Private	Monomers	4	13	8	7,122	54,000	0.1021
Alleles	Total	10	35	22	7,226	46,300	0.0731
Without	Multimers	7	13	-	5,402	30,714	0.0705
Private	Monomers	4	5	-	4,172	23,750	0.0864
Alleles	Total	11	18	-	4,955	28,182	0.0763
TOTAL	Multimers	13	35	14	6,276	35,538	0.0628
	Monomers	8	18	8	5,647	38,875	0.0942
	Total	21	53	22	6,036	36,809	0.0748

the invariant loci, namely, GLO, CA_1 and CA_2 have sample sizes more than 3 S.D.'s below mean sample size and 2 S.D.'s below the lower limit of 4,441. Clearly the small sample sizes at these loci may have affected the detection of rare variants. After excluding these loci, the values of $\hat{k}_t$, $\hat{k}_r$ and $\hat{K}$ are increased by 16.67%.

Considering the islands separately, the estimates of k_t, k_r and K are 0.11, 0.11 and 0.20 respectively for Karkar Island. The respective values for Siassi Islands are 0.12, 0.12 and 0.26 and zero for Manus Island.

Since the calculation of K depends *a priori* on the observed distribution of copies of rare alleles, $\tilde{g}(i)$, and the ratio (f) of sample (n) to effective population size (N), these estimates are inflated by 1.79 and 2.26 times in Karkar and Siassi Islands respectively when compared with the observed value of k_t. This increase is comparable with a similar increase for the estimate of the total sample.

3.2.2.4. Estimation of actual (apparent) and variable effective population sizes.

In the absence of historical records, it is difficult to estimate accurately the actual population size (N) of linguistic groups in Papua New Guinea. However, since the estimates of mutation rate are highly dependent on the estimate

of N, this is discussed in some detail. The Papua New Guinea Bureau of Statistics Census of 1971 reported a total population of 2,435,409 indigenous persons of whom 41.6% were in the reproductive age group of 15-44 years, with a similar proportion (41.5%) ever married. With a minimum of 700 documented language groups (Wurm, 1975a), the maximum estimate of language group effective size is 1,447.

This maximum value of N may be considered a gross overestimate of population size during past generations. Post-Second World War availability of medical care has had a profound impact on mortality, especially that in infants, such that annual population growth currently exceeds 2.3% (van de Kaa, 1971-72). The Census of 1950 (Annual report, 1951) enumerated the total population of Papua New Guinea as 1,440,000, less than 60% of the number observed one generation later. Prior to 1950, many language groups, including those in the Highlands, had hardly been exposed to administrative contact and had not been enumerated. Only a few groups, mostly coastal, had been censused before the Second World War. One such coastal group, included in our sample, is that of Karkar Island. The Island's population increased from 7,300 in 1925 to 9,100 in 1937-9, to 20,068 in 1974 (van de Kaa, 1971-72;

Z'Graggen, 1975a, b). However, such population growth prior to 1939 is exceptional and the corresponding figures for Manus and surrounding islands are 12,900 in 1925, 12,800 in 1937-9 and 24,000 in 1971. Van de Kaa (1971-72) considers that the Papua New Guinea population was stable between 1890 and 1939, partly because there is no evidence to suggest otherwise, but mainly because analysis of the few surveys undertaken at that time show little demographic change.

In the calculations it is assumed that the actual population size during the last five to ten generations more closely approximates the census figures of 1939 - than of 1971, and that the population of 1939 was very close to 50% of that enumerated in 1971.

In Papua New Guinea, estimates of N of language groups also require correction for the extreme variability in language group size. Linguistic groups may comprise less than 100 persons, as in Gorovu in the Ramu Phylum (Z'Graggen, 1971) or more than 150,000 persons as in Enga in the Western Highlands (Wurm, 1975b). By far the majority of language groups have less than 5,000 speakers. For instance, of the 92 languages documented by Z'Graggen (1971) in the Madang Province, only 5% are spoken by more than 5,000 persons and 68% of languages have less than 1,000

speakers. Since the average value of N more closely approximates the harmonic than the simple mean of language group size, the three main linguistic Phyla represented in the Madang Province and analyzed to estimate the ratios of the harmonic means ($\bar{H}$) to simple means ($\bar{N}$). For the Adelbert Range Phylum, $\bar{H}/\bar{N}$ is 35%, for the Ramu Phylum 33% and for the Madang Phylum, 40%. The combined value for 80 languages is 36%.

Therefore, for estimation of the mean number of speakers per language, 50% of 2,435,409 individuals is taken as the total population prior to 1939, distributed amongst 700 languages of varying sizes, with an average of 1,740 speakers. Since the harmonic mean of language group size is 36% of the simple mean, the more appropriate estimate is 626 speakers per language when correction is made for variability in language group size.

The estimation of variance effective population size ($N_e v$) is further modified by the proportion in the reproductive age groups, variability in fertility and deviation of the sex ratio from 1:1. The adult sex ratio was less than unity in the 1971 Census and greater than unity in the previous Census of 1966 (Territory of Papua New Guinea Bureau of Statistics, 1972) so we shall assume the sex ratio in the reproduc-

tive age groups fluctuates around 1:1. The proportion in reproductive age groups is more difficult to estimate accurately, since profound demographic changes make presently observed proportions different from those expected in the past. In 1971, 41.6% of the population was aged between 15 and 44 years, compared with 45.0% in 1966 (Territory of Papua New Guinea Bureau of Statistics, 1972), reflecting the reduced mortality in both infants and the elderly. The Papuan Annual Reports from 1919 to 1937 reported children aged less than 16 years as only 39-40% of the total population and Serjeantson (1970) recorded 49% of the population of two relatively unacculturated PNG language groups aged between 16 and 45 years. It is reasonable to suggest that the proportion in the reproductive age groups in past generations was closer to 49%, the estimate I use in my calculations, than to the 42% presently observed.

Variation in fertility will modify $N_e v$ if the index of variability $(V_{ka}/\bar{k}_a)$ deviates from unity (Crow and Morton, 1955), when $\bar{k}_a$ and V_{ka} are the mean number and variance of surviving offspring. In Papua New Guinea, the index of variability is inflated by factors such as polygyny which was reported by 9% of married males as recently as 1971 (Territory of Papua New Guinea Bureau of Statistics, 1972). Serjeantson (1970) estimated

the index of variability as 1.22 in males from the Yonggom group with 10% polygyny and 2.09 in males from an additional group (Awin) with 28% polygyny. The corresponding values in females were 0.96 and 1.40 in a population with such comparatively low fertility (Serjeantson, 1975) that it may well reflect the demographic structure of most Papua New Guinea groups prior to 1939.

With an average index of variability of 1.4 and 49% of the population in the reproductive age group, the ratio of $N_e v/N$ is 83.7%. The average actual size (N) of language groups in Papua New Guinea is estimated as 49% of 626, or 307 persons and this is the value used in estimating the mean survival time for fresh mutations in Papua New Guinea language groups.

The survey of genetic markers encompassed 55 language groups with a total of 416,215 speakers as shown in table 3.5. In general, it is the language groups with a relatively large number of speakers that have been sampled, so that the average effective size of language groups with genetic data available exceeds slightly the average size of language groups in Papua New Guinea as a whole. Making similar adjustments as before, for rapid population expansion in the last generation, for variation in language size, for variation in fertility and for the

proportion in the reproductive age groups, the total effective population size for the 55 languages in this series is 34,450. I use this value in all my calculations.

The census sizes of the three island populations, namely, Karkar, Manus and Siassi, were about 20,068 (Z'Graggen, 1975a, b), 20,000 (Healey, 1975) and 9,025 (Department of Chief Minister and Department of Administration, 1970-74) respectively in 1973-74. The respective populations stood at 9,110, 13,839 and 4,715 in 1937-39 (van de Kaa, 1971-72) with 50.3%, 62.7% and 59.5% in the adult age groups. After adjusting for 44 years + populations, the values are approximately 41.0%, 53.0% and 49.0% giving values of N as 3,735, 6,805 and 2,310 for Karkar, Manus Island and Siassi Islands respectively. The ratio $N_e v/N$ in Karkar and Siassi Islands is 0.907 and 0.912 respectively.

3.2.2.5. Estimation of $\bar{t}_o$.

The mean survival time for fresh mutations which will ultimately be lost from the population ($\bar{t}_o$) was given by Kimura and Ohta (1969) and Nei (1971). This value is estimated for a Papua New Guinea language group as:

$$\bar{t}_o = 2\ N_e v/N \log_e (2N)$$

$$= 2 \times (0.837) \log_e (2 \times 307)$$

$$= 10.74 \text{ generations}$$

which is different from the estimate given by Neel and Rothman (1978) of 5.70 for Amerindian populations using simulation results. The estimates for Karkar Island and Siassi Islands, were calculated to be 14.92 and 14.12 generations respectively. These estimates are used for generating mutation rates by Kimura and Ohta's method.

3.2.2.6. Results.

The estimation of mutation rates has been carried out using three indirect methods of Kimura and Ohta (1969), Nei (1977) and Rothman and Adams (1978). Table 3.9 shows these estimates for the total Papua New Guinea population. The three estimates of μ by the methods of Kimura and Ohta (1969), Nei (1977) and Rothman and Adams (1978) are 2.83×10^{-6}, 1.44×10^{-6} and 6.58×10^{-6}/locus per generation respectively with a mean value of 3.62×10^{-6}/locus per generation. These estimates range from approximately 28 to 52% of similar estimates obtained for the Australian Aborigines.

The estimates of mutation rate for island populations show a wide range. The value of μ

TABLE 3.9 MUTATION RATES IN PAPUA NEW GUINEA

POPULATION	$\mu \times 10^6$		
	Kimura and Ohta's method	Nei's method	Rothman and Adam's method
Total	2.83	1.44	6.58
Karkar Island	1.79	2.71	5.77
Manus Island	0.00	0.00	0.00
Siassi Islands	4.07	7.56	12.41

in Manus for a set of 14 protein loci is zero. The estimates of μ for Karkar and Siassi Islands are given in table 3.9. The process of estimating the total number of variants in the populations with limited observations, however, is highly unreliable. The estimates of μ obtained in these islands are, however, comparable to similar estimates generated for the Waljbiri tribe in Australian Aborigines.

3.2.2.7. Discussion.

The estimates of mutation rates as obtained from a set of protein loci are affected seriously by a number of factors. Probably the most controversial aspect of these indirect estimates is the estimation of actual population size (N). This is particularly difficult in the Papua New Guinean communities which have recently been undergoing tremendous demographic changes. The impact of recent population expansion can be judged from the high proportion of private polymorphisms with limited geographical distributions. Out of 26 variants detected as many as ten have attained polymorphic proportions in various Papua New Guinean communities, six of them in the highlands, one in Karkar and Siassi and three in both highland and coastal communities. The estimation of N from the present census sizes will, in general, be inflated and it will be appropriate to approach

the subject using a pre-1939 census size.

The role of sample size and subunit size in affecting the detection and introduction of rare variants has been stressed by a number of authors; for example, Nei *et al.* (1976a) and Bhatia (1980) and will be discussed in detail in chapter 4. The mean sample sizes for loci with and without variants are 7,226 and 4,955 respectively. This difference emphasizes the need for a sample size of at least 3,000 even for a set of loci as suggested by Eanes and Koehn (1978), before any attempts are made to generate mutation rates. Similarly, the mean subunit size for loci with these variants is 46,300 daltons compared to 28,180 for the invariant loci. It is thus important to make comparisons of mutation rates among populations only with similar mean sample sizes and mean subunit sizes.

However, the mean number of rare variants/locus per individual (k_t/n) is higher in Australian Aborigines (2.46×10^{-4}) in comparison with Papua New Guinean communites (1.74×10^{-4}). Since the mean number of electromorphs recovered is a logarithmic function of sample size and the distribution of electromorphs is skewed further with sample size increase, it will be appropriate to compute the mean number of rare variants/locus per individual only in terms of actual population size

(N), rather than in terms of sample size (n). The two estimates for Australian Aborigines and Papua New Guineans then become 6.99×10^{-5} and 3.05×10^{-5}, respectively, a difference of 2.29 times.

Nei and Chakraborty (1976) have shown that the mean number of silent alleles, undetectable by electrophoresis, which contribute to an electromorph, is higher in populations with large $N_e\mu$'s than in populations in which this is small. On the basis of this argument, the proportion of mean numbers of silent alleles is likely to be much higher in Papua New Guinean communities than in Australian Aborigines. Since the mean number of private variants/locus (k_t) reflects the incidence of mutation in a population, the ratio of $N\hat{k}_t$ in these two populations, when adjusted for sample sizes, yields a value of 2.66 times more silent alleles in Papua New Guineans than Australian Aborigines. The results at electromorph level (not withstanding the differences in mutation rate at codon level between the two populations) are almost negligible.

Because of these various factors which may affect the mean number of private variants per locus, the indirect estimates of mutation rate will show similar variations. It is not surprising, therefore, that estimates of μ generated from protein data for Papua New Guinea differ about

twofold from estimates generated on a similar scale for Amerindians by Neel and Rothman (1978) and for Australian Aborigines by Bhatia *et al.* (1979). The estimate of μ for a group of tribes in India, however, is lower by more than an order of magnitude compared with these estimates. This may be because of a recent population increase in India and differences in sample size, in the number of loci studied, and in technical methods employed in blood collection, none of which have been taken into consideration by Chakraborty and Roychoudhury (1978).

3.2.3. Mutation rates in Scheduled Tribes from South India.

In the last few years this department has screened a number of Scheduled Tribes in South India for red cell enzyme and serum protein polymorphisms over sets of 12 to 23 loci. The populations studied are: Kadars (Saha *et al.* 1974), Todas, Kurumbas, Irulas and Malayaryans (Saha *et al.* 1976), Kotas (Ghosh *et al.* 1977; Ghosh, 1977a, b), Savaras and Jatapus (Rao *et al.* 1978), Kolams (Ramesh *et al.* 1979), Chenchus (Ramesh *et al.* 1980), Raj Gonds, Pardhans, Koyas, Konda Reddis, Lambadis and Yerukulas (Blake *et al.* 1981), Konda Kammaras, Koyas (second series) (Veerraju *et al.* 1981) and Gadabas (unpublished material).

A number of other laboratories have reported on red cell and serum proteins in Andhra Pradesh tribals (Bernini *et al.* 1970; De Jong *et al.* 1971, 1975; Santachiara-Benerecetti *et al.* 1972a, b; Goud and Rao, 1977, 1980; Papiha *et al.* 1979; Rao *et al.* 1979; Rao and Goud 1979 and Roberts *et al.* 1980) In addition comparative results are available for the non-tribal populations of south India (see Basu, 1978 for a list of references) as well as tribal and non-tribal populations from the adjoining states of Maharashtra, Madhya Pradesh and Orissa, which fall outside the area defined here, and south India. The latter information is valuable, however, in designating private variants.

3.2.3.1. The study population.

The 18 tribal populations included in this study have been divided into three groups on the basis of their geographical proximity and demographic features. Group I comprises nine tribal populations from the northern (Adilabad, Warrangal, Khammam, Srikakulam, Vishakhapatnam, E. Godavari and W. Godavari) districts of Andhra Pradesh; these populations have been grouped together because of their relatively large population sizes, continuous dispersion, positive growth trends in the past 100 years and cultural affiliations with the Scheduled Tribes of central India. In some

cases data from the same, or adjoining, districts have been pooled. Group II includes Chenchus, Lambadis and Yerukulas from Mahabubnagar and Kurnool Districts in southern Andhra Pradesh. They have been grouped together because they were sampled for the same districts but have a discontinuous distribution in restricted pockets over large areas with small effective breeding units. Group III is constituted by six small tribes, all restricted to the Nilgiris and Annamalai Hills of Kerala and Tamil Nadu States. The list of tribes studied is given in table 3.10 and their geographical position in fig. 3.3.

Only populations screened for at least five protein loci have been included in the survey. Table 3.11 gives the number of persons tested for the total 30 red cell and serum protein loci screened. Data generated by the use of non-electrophoretic methods for haemoglobins and glucose-6-phosphate dehydrogenase have not been utilized.

3.2.3.2. The laboratory data.

An allelic variant is considered to be 'private' if it occurs uniquely in only one population (Neel, 1978a). In addition, if any variant allele has a frequency of less than 1% it is considered to be 'rare'. If an electrophoretic variant has been reported in populations from a

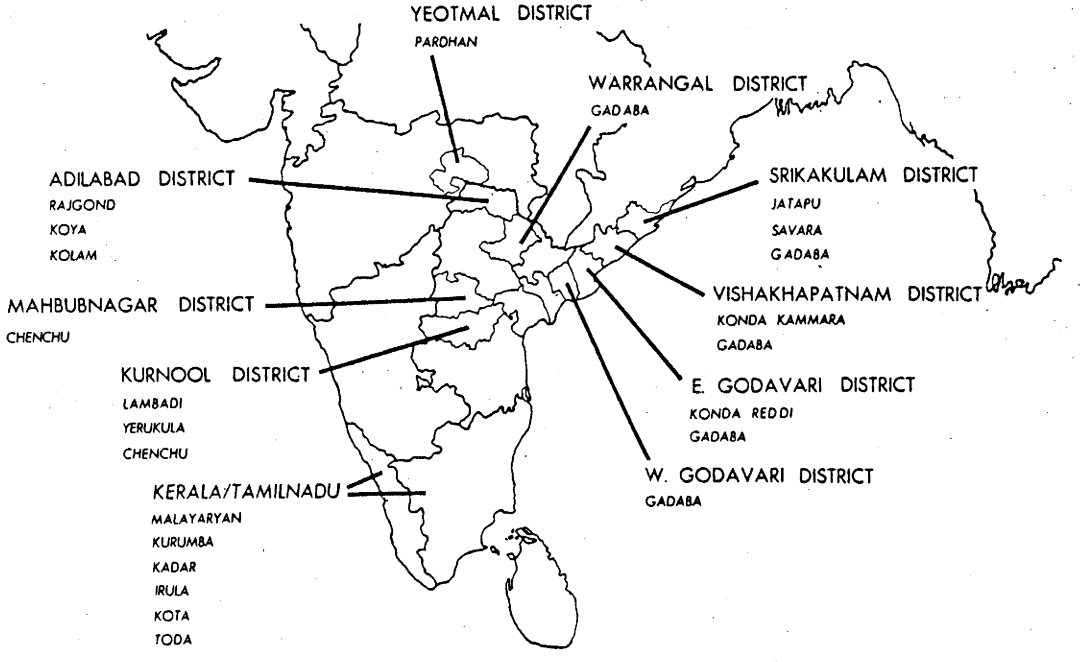

Fig. 3.3 Map of Southern India showing geographical location of various Scheduled Tribes included in the present investigation.

TABLE 3.10 ACTUAL POPULATION SIZE (N) IN SCHEDULED TRIBES OF SOUTH INDIA

Serial number	Scheduled Tribe	District/State of enumeration	Density per km^2 (D)	Total census size	Proportion in 15-44 yrs age group (ℓ)	Actual population size (N)
	GROUP I*					
1.	Savaras	Srikakulam[1]	8.00	40,228	0.439	9,899
2.	Jatapus	Srikakulam[1]	7.61	38,250	0.461	9,811
3.	Kolams	Adilabad[1]	1.62	8,150	0.402	1,836
4.	Koyas	Adilabad[1]	11.56	58,140	0.408	13,297
5.	Raj Gonds	Adilabad[1]	7.21	36,275	0.400	8,136
6.	Pardhans	Yeotmal[2]	2.42	12,171	0.441	3,003
7.	Konda Reddis	E. Godavari[1]	3.45	17,333	0.407	3,954
8.	Konda Kammaras	Vishakhapatnam[1]	1.12	5,619	0.432	1,360
9.	Gadabas	Srikakulam[1] Vishakhapatnam[#]	1.03	5,201	0.434	1,264
	GROUP II*					
10.	Chenchus	Mahabubnagar & Kurnool[1]	-	7,984	0.403	3,217
11.	Lambadis	Kurnool[1]	-	11,704	0.385	4,511
12.	Yerukulas	Kurnool[1]	-	10,650	0.409	4,357
	GROUP III					
13.	Todas	Kerala	-	930	0.447	416
14.	Kurumbas	Kerala/Tamil Nadu	-	4,073	0.506	2,063
15.	Irulas	Kerala/Tamil Nadu	-	103,039	0.445	45,972
16.	Malayaryans	Kerala	-	4,194	0.556	2,331
17.	Kotas	Tamil Nadu	-	1,188	0.454	539
18.	Kadars	Kerala/Tamil Nadu	-	1,418	0.481	681

* The estimates are adjusted to 1921 level.

** In Group I, census size is given as the 'size of the neighbourhood'; for explanations see text.

1 Andhra Pradesh

2 Maharashtra

\# E. Godavari, W. Godavari and Warrangal Districts.

number of localities in the subcontinent (e.g. PHI 2-1), the corresponding allele has not been included in the list of rare variants since its presence in any particular population could be due to intermixture.

In table 3.12 the private and rare variants are indicated in addition to rare variants which are found in more than one population. The latter include especially $Hb\beta^{S}$, LDH_{B}^{Cal-1} and PGD^{C} which are present at low frequency or are absent in some of these tribes. A few others (e.g. PHI^{3}), though rare in general, sometimes achieve polymorphic frequency in one particular population. All of the rare variants in this latter category, of course, are excluded from the calculation of mutation frequency.

Seven private allelic variants are restricted to the populations of this group (table 3.12). Two of these assume polymorphic proportions. PGD^{Gadaba} (1.14%) in Gadabas (unpublished data) and $Hb\alpha^{Koya\ Dora}$ (5.00%) in a sample of Koyas from Polavaram Taluk in West Godavari District (De Jong *et al.* 1975).

Only two private allelic variants, both polymorphic, were detected in group II populations, PHI^{5} and PGM_{2}^{10}. PHI^{5} has been included although a single copy of PHI^{5} has been reported previously from north India (Blake *et al.* 1971). It is unlikely that its presence in the Chenchus is due to admixture.

TABLE 3.11 LIST OF RED CELL ENZYMES, PROTEINS AND SERUM PROTEINS INCLUDED IN THE STUDY OF SCHEDULED TRIBES OF SOUTH INDIA

Serial number	System	Abbreviation	Sample size	Groups (n>1000)	Populations studied*
	Red Cell Enzymes				
1.	Acid phosphatase-1	ACP_1	3,619	I, II	1-18
2.	Adenosine deaminase	ADA	60	-	1
3.	Adenylate kinase-1	AK_1	3,824	I, II	1-7, 9-18
4.	Carbonic anhydrase-1	CA_1	394	-	9,10,12
5.	Carbonic anhydrase-2	CA_2	1,021	-	3,9,10,12,17
6.	Esterase D	EsD	2,023	I	3,4,8-10,12,17
7.	Glucose-6-phosphate dehydrogenase	G-6-PD	213	-	18
8.	Glutamic oxaloacetic acid transaminase	GOT	359	-	1-3
9.	Glyoxalase-1	GLO_1	756	-	4,8,17
10.	Isocitrate dehydrogenase	ICD	1,497	II	3,10,12-18
11.	Lactate dehydrogenase-A	LDH_A	3,575	I,II	1-18
12.	Lactate dehydrogenase-B	LDH_B	3,575	I,II	1-18
13.	Malate dehydrogenase-2	MDH_2	3,646	I,II	1-18
14.	Nucleoside phosphorylase	Np	580	-	3,13-16,18
15.	Peptidase A	Pep A	1,012	-	1-3,13-16,18
16.	Peptidase B	Pep B	1,012	-	1-3,13-16,18
17.	Peptidase D	Pep D	693	-	1-3,13-16
18	Phosphogluconate dehydrogenase	PGD	3,610	I,II	1-18
19.	Phosphoglycerate kinase	PGK	2,804	I,II	1-3,9,10,12-18
20.	Phosphoglucomutase-1	PGM_1	4,014	I,II	1-18
21.	Phosphoglucomutase-2	PGM_2	4,022	I,II	1-18
22.	Phosphohexose isomerase	PHI	3,644	I,II	1-18
23.	Superoxide dismutase	SOD_A	3,644	I,II	1-18
	Red Cell Proteins				
24.	Hemoglobin-α	Hb-α	4,159	I,II	1-18
25.	Hemoglobib-ß	Hb-ß	4,159	I,II	1-18
	Serum Proteins				
26.	Albumin	Alb	1,829	-	4-6,10,11,13-16,18
27.	Caeruloplasmin	Cp	2,208	II	4-6,11,13-18
28	Group specific component	Gc	1,006	-	4-6,11,14,15
29.	Haptoglobin	Hp	2,477	II	4-6,10,11,13-18
30.	Transferrin	Tf	2,374	II	4-6,10,11,13-18

* The serial numbers of the populations are given in Table I.

TABLE 3.12 RARE VARIANTS AND PRIVATE POLYMORPHISMS IN SCHEDULED TRIBES OF SOUTH INDIA

Population	Single locus determinations	Allelic variants (percent frequency)	References (See notes)
Group I			
Savaras	2,150	$Hb\beta^S$ (0.76)	13
Jatapus	2,664	$Hb\beta^S$ (0.64), PHI^3 (0.65)	13
Kolams	3,818	PHI^2 (0.23), PGM_1^7 (0.47)	12
Koyas	7,897	$Hb\alpha^{Koya\ Dora}$ (0.99)*, $Hb\alpha^{Rampa}$ (0.05)*, PHI^2 (0.19), PHI^3 (1.36), $Alb^{Koya\ Dora}$ (0.09)*, $Tf^{D\ Chi}$ (2.92), PGM_1^4 (0.11)*, PGM_1^6 (0.11)	1-4,8,14,15, 18-20
Raj Gonds	3,559	PHI^3 (0.37), PGD^C (0.75), $Tf^{D\ Gond}$ (0.70)*, $Tf^{D\ Chi}$ (1.04)	2,8,14,15
Pardhans	1,952	-	2,8,14,15
Konda Reddis	1,657	ACP_1^C (0.55)	2-4,18-19
Konda Kammaras	1,429	PHI^2 (0.45), PGD^C (0.93)	20
Gadabas	13,035	PHI^6 (0.45)*, PGD^{Gadaba} (1.14)*, PGM_1^7 (0.10)	Unpublished
Group II			
Chenchus	3,521	$Hb\beta^S$ (0.26), PHI^2 (0.25), PHI^5 (2.71)*, LDH_B^{Cal-1} (1.48), Tf^D (0.35), PGM_1^7 (0.25)	11
Lambadis	1,674	$Tf^{D\ Chi}$ (0.32)	2,8,14,15
Yerukulas	569	PGM_1^7 (7.69), PGM_2^{10} (1.25)*	2
Group III			
Todas	2,303	$Hb\beta^S$ (0.51), PGD^C (0.51), LDH_A^{Toda} (0.51)*	10,17
Kurumbas	1,049	LDH_B^{Cal-1} (3.49)	9,10,17
Irulas	3,765	LDH_B^{Cal-1} (0.29)	9,10,17
Malayaryans	1,280	$PGM_1^{6\ Mal}$ (7.63)*	17
Kotas	10,816	-	5-7
Kadars	4,671	$Hb\beta^S$ (0.47), LDH_B^{Cal-1} (1.64), PGD^K (4.24)* ACP_1^C (0.48), PGM_1^{6K} (2.11)*, $Pep\ B^K$ (0.23)*	17

* Private (rare/polymorphic) allele.

Notes: 1. Bernini *et al.* (1970); 2. Blake *et al.* (1981); 3. De Jong *et al.* (1971); 4. De Jong *et al.* (1975); 5. Ghosh (1977a); 6. Ghosh (1977b); 7. Ghosh *et al.* (1977); 8. Goud and Rao (1977); 9. Kirk *et al.* (1963); 10. Kirk *et al.* (1962); 11. Ramesh *et al.* (1980); 12. Ramesh *et al.* (1979); 13. Rao *et al.* (1978); 14. Rao and Goud (1979); 15. Rao *et al.* (1979); 16. Saha *et al.* (1974); 17. Saha *et al.* (1976); 18. Santachiara-Benerecetti *et al.* (1972a); 19. Santachiara-Benerecetti *et al.* (1972b); 20. Veerraju *et al.* (1981).

A number of private (rare as well as polymorphic) variants have been observed in three of the six tribal populations in this group. The private polymorphisms are PGM_1^{6K}, PGD^{Kadar} and $PGM_1^{6\ Mal}$, the former two in Kadars and the latter in Malayaryans, and two rare variants $PEP\ B^{K}$ in Kadars and LDH_A^{Toda} in Todas.

3.2.3.3. Estimation of k_t, k_r and K.

Tables 3.13 and 3.14 show the various estimates of the mean number of variants/per locus, observed in the sample ($\hat{k}_t$ and $\hat{k}_r$) and estimated from the sample number of alleles for the total population of alleles. $\tilde{g}(j)$, were obtained from the observed distribution of copies of various rare variants in the pooled data. The estimates of b and c obtained from the results of the demogenetic survey of Kota by Ghosh (1976) were utilized in computing P_{j1} by the equation:

$$P_{j1} = j\,\hat{b}\,\left(1-\frac{\hat{b}}{1-\hat{c}}\right)^{j-1} \qquad (3.2)$$

It may be noticed that no values are obtained for Savaras, Jatapus, Kolams, Pardhans, Konda Reddis, Konda Kammaras, Lambadis, Kurumbas, Irulas and Kotas. In addition, no estimate of k_r was obtained for Chenchus, Yerukulas and Malayaryans (table 3.13).

TABLE 3.13 SUMMARY OF VARIOUS GENETIC PARAMETERS AND SAMPLE SIZE IN 18 SCHEDULED TRIBES FROM SOUTH INDIA

Serial number	Population	Number of cistrons	Mean per locus determinations (n_i)	Private rare variants	$\hat{t}_o$**	Variants/locus		
						$\hat{k}_t$	$\hat{k}_r$	$\hat{K}$
	Group I*							
1.	Savaras	18	179.44 ± 6.31	0(0)	16.205	-	-	-
2.	Jatapus	17	156.71 ± 0.19	0(0)	16.202	-	-	-
3.	Kolams	21	181.61 ± 10.21	0(0)	13.446	-	-	-
4.	Koyas	19	415.63 ± 57.75	4(4)	16.689	0.2105	0.2105	0.7854
5.	Raj Gonds	17	209.75 ± 25.65	1(1)	15.884	0.0588	0.0588	0.2446
6.	Pardhans	17	114.82 ± 9.98	0(0)	14.251	-	-	-
7.	Konda Reddis	12	138.08 ± 24.12	0(0)	14.702	-	-	-
8.	Konda Kammaras	13	109.92 ± 0.08	0(0)	12.954	-	-	-
9.	Gadabas	16	814.69 ± 86.13	2(1)	12.834	0.1250	0.0625	0.1448
	AVERAGE							
	Group II*							
10.	Chenchus	20	176.05 ± 6.84	1(0)	14.364	0.0500	-	0.1378
11.	Lambadis	17	98.47 ± 14.11	0(0)	14.918	-	-	-
12.	Yerukulas	17	33.47 ± 2.42	1(0)	14.861	0.588	-	0.4541
	Group III							
13.	Todas	22	104.68 ± 5.26	1(1)	8.741	0.0455	0.0455	0.0688
14.	Kurumbas	23	45.61 ± 2.81	0(0)	10.823	-	-	-
15.	Irulas	23	163.70 ± 8.26	0(0)	14.858	-	-	-
16.	Malayaryans	22	58.18 ± 0.52	1(0)	10.981	0.0455	-	0.1926
17.	Kotas	20	540.80 ± 6.93	0(0)	9.078	-	-	-
18.	Kadars	22	212.32 ± 0.36	3(1)	9.382	0.1364	0.0455	0.1941
	AVERAGE							

* Actual population size adjusted to 1921 census numbers.

** The value of $N_e v/N$ is 0.819, 0.819 and 0.650 in Group I, Group II and Group III respectively.

3.2.3.4. Estimation of actual population size (N) and variance effective population size ($N_e v$).

The actual (apparent) population size N, (effective population size in Nei's (1977) and Bhatia *et al.* (1979)'s terminology) is an important parameter and is normally equated to the number of breeding individuals given as the proportion of the population in the age group 15-44 years. In populations with cyclic changes in population size over the past few generations, Wright (1931) has recommended the use of harmonic mean.

To accommodate the role of the isolation by distance in the large, continuously dispersed populations of group I, the value of N is estimated as the 'size of neighbourhood', following Wright (1946), as

$$N = 4\pi\sigma^2 D\lambda$$

where σ^2 is the variance of migrational distances, D is the population density and λ the proportion of the reproductive age group (15-44 years) in the population. Pingle (1975)'s data yield the variance of marital distance in Adilabad tribes to be approximately 400 km^2. Majumdar (1977) has, however, given much smaller values for marital distances for the Andhra populations as a whole. The individual values of D and λ estimated from

age and sex tables of the Andhra tribes (Census of India, 1971) are shown in table 3.10.

The effective breeding unit in the discontinuously distributed populations of group II is taken to be the administrative district where the samples have been collected (table 3.10). Total census sizes have been utilized in group III populations for estimating N.

Except Chenchus, who have increased marginally over their 1911 numbers, the populations of groups I and II have been adjusted for population increase since 1881. Taking 1921 census as the base level, which is quite close to the harmonic mean size since 1881, I calculate the new estimates of N to be 0.560 of their 1971 numbers. No such adjustment is necessary for group III tribes, which have only recently built up their original numbers to 1881 levels after a decline in the early decades of this century.

The value of the variance effective number $N_e v$, is given (Crow and Morton, 1955; Crow and Kimura, 1972) as

$$N_e v = 2N/[(1-F)+(1+F)V_k/\bar{k}]$$

where F is a measure of departure from Hardy Weinberg proportions, taken formally equivalent to the inbreeding coefficient and $\bar{k}$ and V_k are mean and variance respectively of the progeny size

surviving to adulthood. $V_k/\bar{k}$ also defines the index of variability. At birth and adulthood the mean and variance will be defined by $\bar{k}_b$ and V_{kb}, $\bar{k}_a$ and V_{ka} respectively.

Murty and Ramesh (1979) and Ghosh (1970) have provided the estimates of $\bar{k}_b$ and V_{kb} for post reproductive age women and also the index of mortality I_m (Crow, 1958), for Adilabad tribes and Kotas respectively. The index of variability at adulthood ($V_{ka}/\bar{k}_a$) is recalculated using the formulae

$$P_s = 1/(1+I_m)$$

and $$\frac{V_{ka}}{\bar{k}_a} = 1 + P_s\left[\frac{V_{kb}}{\bar{k}_b} - 1\right]$$

where P_s is the probability of survival to adulthood and the subscripts a and b refer to the values at adulthood and at birth respectively. The estimated values of this index in Adilabad tribes and Kotas is 1.43 and 1.95 respectively. Basu's (1972) data yield a value of 2.03 for Kotas. The high value in Kotas is attributed to a large proportion of nulliparous women in the 45 + years age group. Similar demographic trends are seen in Irulas (Basu 1967). Since no published results are available on progeny size for men and women separately, adjustments for variation due to polygamy are not made in these calculations.

The value of F obtained from pedigree data on Andhra tribes and Kotas is 0.030 (Veerraju, 1978) and 0.040 (Ghosh, 1972; 1976) respectively. Inserting these values of F and the respective estimates of $V_{ka}/\bar{k}_a$ in the equation, $N_e v/N$ becomes 0.819 and 0.650 in Adnhra tribes and Kotas respectively. The former value is used for computing $\bar{t}_o$ in individual populations of groups I and II, and the latter for the populations of group III.

3.2.3.5. Estimation of $\bar{t}_o$.

Another important parameter used in estimating mutation rates by Kimura and Ohta's method is the expected number of generations a mutant survives prior to extinction $\bar{t}_o$. The values of $\bar{t}_o$ for various tribes are shown in table 3.12. For estimating mutation rate in combined tribes, group I, II and III, the mean value of $\bar{t}_o$ is computed over nine, three and six tribes respectively. For the total population, the average of $\bar{t}_o$ values over 18 tribes is used. These values of $\bar{t}_o$ are given in table 3.14.

3.2.3.6. Results.

The results on the indirect estimation of mutation rate obtained at three levels of population organisation, i.e. at individual population, individual group and all tribes level are presented in tables 3.15 and 3.16. The estimators used are $\hat{\mu}_{K-O}$,

TABLE 3.14 SUMMARY OF VARIOUS GENETIC PARAMETERS AND NUMBER OF SINGLE LOCUS DETERMINATIONS FOR VARIOUS GROUPS OF SCHEDULED TRIBES OF SOUTH INDIA

Population group	Number of loci	Single locus determi-nations	Private (rare) variants	N (15-44yrs)	$\hat{t}_o$	Variants/locus		
						$\hat{k}_t$	$\hat{k}_r$	$\hat{K}$
				ALL LOCI				
Group I	29	38,161	7(7)	52,560	14.796	0.2414	0.2414	1.0220
Group II	22	5,764	2(1)	12,085	14.714	0.0909	0.0455	0.4166
Group III	27	23,884	5(5)	52,002	10.644	0.1852	0.1852	0.9712
				LOCI WITH n>1,000 TESTS				
Group I	14	30,868	5(5)	52,560	14.796	0.3517	0.3517	1.1165
Group III	17	19,435	4(4)	52,002	10.644	0.2353	0.2353	1.0700
				TOTAL POOLED DATA				
Total loci	30	67,809	14(14)	116,089	13.398	0.4667	0.4667	2.2833
Loci with n>1,000	23	64,754	14(14)	116,089	13.398	0.6087	0.6087	2.6260

TABLE 3.15 ESTIMATES OF MUTATION RATE IN 18 SCHEDULED TRIBES OF SOUTH INDIA

		$\mu \times 10^6$		
		$\hat{\mu}_{K-O}$	$\hat{\mu}_{NEI}$	$\hat{\mu}_{R-A}$
	Group I			
1.	Savaras	-	-	-
2.	Jatapus	-	-	-
3.	Kolams	-	-	-
4.	Koyas	1.7	1.8	7.3
5.	Raj Gonds	0.9	1.2	3.7
6.	Pardhans	-	-	-
7.	Konda Reddis	-	-	-
8.	Konda Kammaras	-	-	-
9.	Gadabas	4.4	4.4	14.1
	Average	0.7	0.8	2.8
	Group II			
10.	Chenchus	1.4	-	5.3
11.	Lambadis	-	-	-
12.	Yerukulas	3.5	**	12.9
	Average	1.6	-	6.0
	Group III			
13.	Todas	9.4	37.0	20.4
14.	Kurumbas	-	-	-
15.	Irulas	-	-	-
16.	Malayaryans	3.7	-	10.2
17.	Kotas	-	-	-
18.	Kadars	15.1	11.5	35.3
	Average	4.7	8.0	10.9
	Overall Average	2.2	3.1	6.0

** No estimate possible for sample size < 1/q

TABLE 3.16 ESTIMATES OF MUTATION RATE IN VARIOUS GROUPS OF SCHEDULED TRIBES OF SOUTH INDIA

	$\hat{\mu} \times 10^6$		
	$\hat{\mu}_{K-O}$	$\hat{\mu}_{NEI}$	$\hat{\mu}_{R-A}$
ALL LOCI			
Group I	0.657	0.351	2.410
Group II	1.171	0.568	4.273
Group III	0.877	0.310	2.315
AVERAGE	0.902	0.410	2.999
LOCI WITH n>1000 TESTS			
Group I	0.718	0.442	2.633
Group III	0.966	0.361	2.550
AVERAGE	0.842	0.401	2.592
TOTAL POOLED DATA			
ALL LOCI	0.734	0.264	2.438
LOCI WITH n>1,000	0.844	0.325	2.804

$\hat{\mu}_{NEI}$, $\hat{\mu}_{R-A}$ as given by Kimura and Ohta (1969), Nei (1977) and Rothman and Adams (1978) respectively.

At individual population level the estimates of μ show wide variability (table 3.15). Even for non-null results the values differ by more than an order of magnitude.

The estimates of μ in group III populations are on an average higher than those obtained for groups I and II populations. The unweighted average of these 18 individual population estimates is 2.26×10^{-6}/locus per generation, 3.12×10^{-6}/locus per generation, 6.08×10^{-6}/locus per generation by the methods of Kimura and Ohta (1969), Nei (1977) and Rothman and Adams (1978) respectively (table 3.15). These estimates, however, entail large standard errors (SE) which may be contributed by fluctuations in the estimates of k_t, $\hat{k}_r$ and $\hat{K}$ as also the errors associated with the estimation of N.

The estimates of μ at individual group level, however, do not show much variability. The unweighted averages of three individual group estimates are 0.902×10^{-6}, 0.410×10^{-6} and 2.999×10^{-6}/locus per generation by the procedures of Kimura and Ohta (1969), Nei (1977) and Rothman and Adams (1978) respectively (table 3.16).

The pooled data, over all the 18 populations, yields much smaller estimates of μ. The values of $\hat{\mu}_{K-0}$, $\hat{\mu}_{NEI}$ and $\hat{\mu}_{R-A}$ are 0.734×10^{-6}/locus per generation, 0.264×10^{-6}/locus per generation and 2.438×10^{-6}/locus per generation respectively.

Large standard errors (SE) are associated with the estimates of $\hat{k}_t$, $\hat{k}_r$ and $\hat{K}$ which are largely due to fluctuations in the sample size which affects the recovery of rare alleles seriously and the variability in the mutation rate over loci on account of subunit size (MW) variations. The variability in one sample size of loci tested for group I is 60-2,589, for group II is 113-1,327 and for the total pooled data is 60-4,159. Since some loci were tested on a relatively small number of individuals I have estimated new values of μ only for those loci which have been tested for at least 1,000 individuals. The new estimates of μ_{K-0} for groups 1 and III are 7.15×10^{-6}/locus per generation and 0.966×10^{-6}/ per locus per generation with a mean value of 0.842×10^{-6}/locus per generation. Similar estimates of μ_{NEI} and μ_{R-A} are shown in table 3.16. These estimates are slightly higher than the earlier estimates for groups I and II by the three methods.

The estimates of μ_{K-0}, μ_{NEI} and μ_{R-A} for pooled date over 23 loci (n>1,000) are 0.844×10^{-6}/ locus per generation, 0.325×10^{-6}/locus per gener-

ation and 2.804×10^{-6}/locus per generation respectively (table 3.16). The differences of these estimates from those obtained previously are only marginal.

3.2.3.7. Discussion.

The indirect estimates of mutation rate generated on the Scheduled Tribes from south India are clearly outside the range of similar estimates when comparisons are made with those obtained at similar levels of population organization on Amerindians (Neel and Rothman, 1978), Australian Aborigines (Bhatia *et al.* 1979) and Papua New Guineans (Bhatia *et al.* 1981). At individual tribe level, the unweighted average of μ for tribes in India is about one fourth of the unweighted average for Amerindians. Similarly, the results on the pooled population of all tribes from south India are considerably lower than similar estimates on the Australian Aborigines and the Papua New Guineans. The present results, although based on a much wider data base, in fact, confirm the apprehensions of Chakraborty and Roychoudhury (1978) regarding the use of data on moderately acculturated Indian tribes for the estimation of mutation rate, although the possibility of regional/ethnic differences in mutation rate exists (Neel *et al.* 1980b).

One of the factors which affect these estimates on Indian tribes seriously are the conservative procedures employed in designating a private variant. In the Indian context, where a large number of communities live together in the same area, identities between electromorphs suggest common descent and fresh mutations are private by default only. Considerable under-estimation of this sort of data makes the genetic interpretation of these results difficult.

Another serious source of error in utilizing the indirect procedures is the use of the parameter N, the actual size of the population. In continuously distributed, large sized population groups, the effective size of the breeding unit estimate by the methods of Wright (1946) or Bhatia *et al.* (1981), suitability of the approach notwithstanding, is only approximate and tends to err on the higher side. In the absence of hard data on the historic demography of these pre-literate societies, analogous dilemmas of a more temporal nature are faced. In addition, the extrapolation of the current demographic compositions to a time-specific constancy is disputable. The much lower estimates of μ in the Papua New Guineans and the Indian tribes may be attributed partly to these over-estimations of the expected harmonic values of N.

The use of electrophoretic data, as analysed by the standard methods, clearly defines only a subset of total mutational events occurring at a given cistron and thus any estimates obtained by these approaches must be adjusted for these underestimations. In addition to about two thirds of the aminoacid substitutions which lead to no charge change (Shaw, 1965; Nei and Chakraborty, 1973; Marshall and Brown, 1975), a large fraction, depending upon the distribution/density of the population and the relative frequencies of the electromorphs, of electrophoretically detectable substitutions is lost due to coalescence with other electromorphs (Nei and Chakraborty 1976; Chakraborty and Nei, 1976; Takahata, 1980). The effect of the latter is correlated with the population size. For presumably similar neutral mutation rates over similar sets of protein loci, Bhatia *et al.* (1981) found 2.66 times more silent alleles in the numerically stronger (and more densely distributed) Papua New Guinean communities, than in the thinly spread, small sized group of the Australian Aborigines. Because of the undefinable nature of these population sizes no adjustments have been made on this accord, though the present estimates may only be 20 to 40 per cent of the real values.

Another class of mutations omitted in these calculations is null mutations. Although the biochemical nature of these mutations may range from a single aminoacid substituion in a polypeptide to a total loss of polypeptide and, in theory, may result from mutations either in structural or regulatory genes (Neel, 1978a), the ratio of μ_{null} to $\mu_{variant}$ is known to range from 2-6fold (Mukai and Cockerham, 1977; Voelker *et al.* 1980a). Arthur *et al.* (1975) and Nelson and Harris (1975) have reported more than 12 fold more null mutations in experiments on mutagenised human cultured cells. Since it is now possible to distinguish between structural and regulatory mutations (Siciliano *et al.* 1978) the proportion of null mutations at structural loci can be estimated. Although we do not introduce any correction for this factor here, it may be noted that such adjustments will raise the estimates considerably, especially on large sized populations.

The procedures for calculating indirect estimates of μ from protein data have now been extended to non-human species by McCommas and Chakraborty (1980). In *Bunodosoma cavernata* they have estimated the $\hat{\mu}$ to be 6.3×10^{-7} to 6.3×10^{-8}/ locus per generation for population sizes of 10^6 to 10^7 individuals. These results, along with my results of Indian and Papua New Guinean populations, indicate that very low values of $\hat{\mu}$ are generated

from protein data by using indirect procedures, specially if the population sizes are large.

The results on direct estimates of mutation rates from protein data on *Drosophila* and man are now available. Mukai and Cockerham (1977) and Voelker *et al.* (1980b) reported the frequency of band morph mutations in *Drosophila melanogaster* as 1.81×10^{-6} and 1.28×10^{-6}/locus per generation respectively. Dubinin and Altukhov (1979) and Neel *et al.* (1980b) have given these estimates in human populations as 6×10^{-5} and 0.34×10^{-5} locus per generation in Russians and Japanese respectively. The results on human populations are, however, difficult to evaluate since more than 522,119 determinations in English (Harris *et al.* 1974), Amerindians (Neel *et al.* 1980a) and Japanese (Neel *et al.* 1980b) have failed to identify a single confirmed instance of spontaneous mutation, although the possibility of detecting much common null mutations also exists. If anything, these results only indicate that the differences in mutation rates between moderately acculturated, comparatively large sized, Scheduled Tribe groups of south India, and the other non-tribal communities, may not be very large. Bhatia (1981) has also indicated that the range of inter-populational estimates of relative electromorph mutation rates is much lower than the range of

their effective population sizes, indicating that mutability differences among human populations, both civilized and primitive, if any, are only marginal.

3.2.4. Estimates of mutation rate in hunter-gatherers of central and southern Africa.

3.2.4.1. The study population.

The present section analyzes data from a number of sources for several hunter-gatherer populations in central and southern Africa. These may be classified into two broad groups: Pygmies from the equatorial forests of central Africa and more heterogeneous groups of peoples who, with one exception, speak "click" or Khoisan languages. The best known of the Khoisan-speakers form a morphologically distinct group which is referred to as Khoisan and this, in turn, may be sub-divided into the Khoikhoi and San.

Of all the people extant in Africa, Pygmies and Khoisan-speakers have the greatest claim to antiquity of residence (Stow, 1905; Vergnes *et al.* 1979). Before the permeation of their territory by Bantu-speaking Negroids from the north and east and Caucasians from the south, Pygmies dominated the equatorial Africa while speakers of Khoisan languages were spread all across the South African subcontinent (Coon, 1965; Hiernaux, 1976).

Those who could assimilated into the more settled communities, whereas others who chose to retreat are still living a lot of their original life-style of hunting and gathering in isolated pockets over central and southern Africa.

Characterized by short stature and distinct economic structure, Pygmies have also retained distinct social customs. However, Pygmies have lost their original language after contact with Negroid farmers (Cavalli-Sforza, 1972). Although the genetic profile of the neighbouring farmers has been modified by the assimilation of Pygmy genes, Pygmies possess a set of genetic variants which distinguish them clearly from the neighbouring farmers in terms of gene frequencies and private variants (Sahtachiara-Benerecetti *et al.* 1980; Vergnes *et al.* 1979 and Beretta *et al.* 1977).

Recently, Jenkins *et al.* (1978), Nurse *et al.* (1978) and Nei and Roychoudhury (1981) have shown through genetic distance analyses that the Khoisans are distinct from other Negroid groups. In addition, Khoisans exhibit a number of rare variants and private polymorphisms of plasma proteins (Jenkins and Steinberg, 1966; Steinberg *et al.* 1975), the blood groups and erythrocyte enzymes (Jenkins, 1974; Jenkins *et al.* 1971; 1975; Jenkins and Nurse 1976; Nurse *et al.* 1977 and Nurse and Jenkins, 1977a) and the tissue antigens

(Botha *et al.* 1973; Nurse *et al.* 1975), which indicate a considerable Khoisan divergence from Negroids. The two distinctive sub-divisions of the Khoisan group, namely Khoikhoi and San, are known to differ from each other and only in minor biological traits (Nurse 1977).

Many of the other Khoisan-speaking groups are morphologically similar to Bantu-speaking Negroids. Their genetic distinction, however, both from the Khoisan and Bantu-speakers has been demonstrated recently by the studies of Godber *et al.* (1976) on Sandawe of Tanzania, Nurse *et al.* (1976) and Knusmann and Knusmann (1969-70) on Dama of Namibia, Nurse and Jenkins (1977b) on Kwengo of Western Caprivizipfel region of Namibia and Nurse and Jenkins (1977b) on Danisan of the Republic of Botswana.

One additional hunter-gatherer group is included in the present analysis, namely the Kgalgadi of Botswana. The Kgalgadi are morphologically similar to Bantu-speaking Negroids and they also speak a Bantu language.

Several workers have documented the fact that when hunter-gatherers are brought into contact with settled agriculturalists there is emigration from the former to the latter, but not *vice-versa* (Cavalli-Sforza *et al.* 1969; Bodmer and Bodmer, 1970; Neyman, 1978). For

the population geneticist it may be considered that the remaining hunter-gatherer isolates are representative of their original gene pools. Assuming that these populations retain private variants of long-standing they can provide valuable data for the indirect estimation of mutation rates.

3.2.4.2. Estimation of actual (N) and variance effective population size ($N_e v$).

3.2.4.2.1. San.

Some 300 years ago, the San people covered the whole of Africa south of the Zambezi river, numbering about 150-300,000 people (Lee, 1976). The systematic extermination campaigns organized by Dutch colonists (Moodie, 1840-42; Marks, 1972) and subjugation by and assimilation with, other ethnic groups (Marais, 1939) led to a precipitous decline in San numbers. While the eclipse of the Cape San was total, Kalahari San gave way to settled agriculturist and pastoralist communities on the Kalahari fringes only to a limited extent. According to Lee (1979), there are approximately 40,500 living San, of which 88.64% are in Republic of Botswana and Namibia. Of the other, about 4,000 live in Angola and a few hundred each in Zambia and Zimbabwe.

The estimates of San numbers given by Lee (1979) are conservative, albeit more reliable, in comparison with the numbers enumerated by other workers

(Silberbauer, 1965; Nurse, 1977; O'Callaghan, 1977; Esterman, 1956 and Guerreiro, 1968). Exclusion of River Basarwa, Kwengo and Danisan, whose genetic profile is non-Khoisanoid (Nurse and Jenkins, 1977a, b) from Lee's estimates reduces the population numbers further to 29,500. Various San groups included in the study, along with their population numbers are listed in table 3.17.

The microdemographic survey of Dobe!Kung San by Howell (1979) reveals 49.56% of the population in the age group 15-44 years (λ). The simulated values of the age composition of Dobe!Kung based on their mortality and fertility schedules fit well to the stable population model "West 5" of Coale and Demeny (1966) which gives the value of $\hat{\lambda}$ as 45.82% (Howell, 1979). It will be more appropriate to use this latter value of $\hat{\lambda}$, as recommended by Howell (1976 , 1979) since the use of a demographic survey from a limited region to specify the global demographic configuration of the San population is unjustified especially when demographic variation amongst various San groups is known to exist (Silberbauer, 1965; Harpending, 1976; Marshall, 1976 and Lee, 1979).

Delineation of the actual breeding group in San is, however, difficult. Evidence from genetic polymorphisms indicates that gene exchange among San does transcend the band and linguistic boundar-

TABLE 3.17 LINGUISTIC GROUPINGS, APPROXIMATE POPULATION SIZE AND LOCATION OF VARIOUS SAN POPULATIONS INCLUDED IN THE STUDY (AFTER LEE, 1979)

	Language (cluster/ group)	Population[1] size (approx.)	Distri-[2] bution	Study Group	Location[3] (Lat. S°, Long E°)	Reference
A.	Tshu-Khwe					
1.	Hei//om	2,000	c	Hei//om	20.5,16.0	Nurse *et al.*, 1977
2.	Naron	6,900	b,c	Nharo	21.5,21.5	-ditto-
3.	/Gwi&//Gana	3,000	b	G/wi & G//ana	23.0,24.0	Jenkins *et al.*,1975
4.	River Basaswa	1,000	b	not studied	17.5,22.5	-ditto-
5.	Kwengo	2,000	a,b,c,d	Kwengo	17.0,23.5	Nurse and Jenkins, 1977b
6.	Danisan	8,000	b,e	Denasena	20.0,25.5	Nurse and Jenkins, 1977a
	Subtotal	22,900				
B.	Southern San					
7.	!Xõ	2,000	b,c	!Xõ&≠huã:	{24.0,22.0 / 25.0,21.5}	Nurse and Jenkins, 1977a
8.	N/huki	~ 100	c,d	G!aokx'ate	-	-ditto-
9.	//Xegwi	-	d	not studied	26.5,30.0	-ditto-
	Subtotal	2,100				
C.	Northern San					
10.	!Kung	6,500	a,b,c	G!ag!ai	18.5,21.0	Nurse *et al.*,1977
11.	≠Dau//keisi	3,000	b,c	//au//en	21.0,20.5	-ditto-
12.	Zhu/twasi	6,000	b,c	Tsumkwe!Kung[4,5] Saman!aika!Kung	19.5,21.5	-ditto-
	Subtotal	15,500				
	Total	40,500				

Note: 1. After Lee (1979).

2. See (1) Also a = Angola; b = Botswana; c = Namibia; d = Zambia; e = Zimbabwe.

3. After Nurse (1972).

4. Tsumkwe is also called Chum!kwe.

5. Also include Dobe!Kung, /du/da !Kung and /ai/ai(or/Xai/Xai)!Kung.

ies (Harpending and Jenkins, 1971; Harpending, 1976 and Nurse and Jenkins, 1977a) although the rate of group endogamy is very high (Marshall, 1976; Harpending and Jenkins, 1971).

A group of bands pooling resources during the lean periods constitutes effectively an 80% endogamous panmictic unit of Adams and Kasakoff (1975) among the San (Howell, 1979; Marshall, 1976). The harmonic mean, of the census size for eight such groups given by Harpending (1976), is ~289 individuals which for $\hat{\lambda}$ = 0.4582 gives the average size of the breeding group in San as 132 individuals.

An alternative estimate of N is given by the panmictic circle approach following Wright (1946) (for formulation see equation 3.3). Harpending (1976) estimates the variance of matrimonial distances in San to be 4142.21km^2 while the population density (D) of !Kung is obtained to be one person for 3.7km^2 [15,500 !Kung (Lee, 1979) spread over an area of 60,000km^2 (Lee, 1976)]. For $\hat{\lambda}$ = 0.4582, the size of the panmictic circle is obtained to be ~6,154 individuals. This estimate of N, however, assumes no migration at the periphery of the panmictic circle.

Variance effective size ($N_e v$) is the size of a population, with discrete generations and binomial sampling of gametes, whose sampling variance is equivalent to the variance associated with gene

frequency change (Crow and Kimura, 1972). For populations with variance in progeny size, this is given by Crow and Kimura (1972) as

$$N_e v = \frac{2N-(k_a/2)}{1+(V_{ka}/\bar{k}_a)} \qquad (3.5)$$

where $\bar{k}_a$ and V_{ka} are the mean and variance of progeny size surviving to adulthood, respectively. For large N, the numerator in (3.5) is approximated to 2N.

Data on mean and variance of progeny size at birth for !Kung men and women have been provided by Harpending (1976) and Howell (1976, 1979). After correcting for the phase of survival from birth to adulthood, we get an average estimate of the ratio $N_e v/N$ as 0.898 (see table 3.18) using the equation

$$\frac{N_e v}{N} = \frac{2}{2+P_s\left[\frac{V_{kb}}{\bar{k}_b} - 1\right]} \qquad (3.6)$$

where P_s is the probability of survival to adulthood and $\bar{k}_b$ and V_{kb} are the mean and variance of progeny size at birth respectively.

However, the use of progeny data only from men and women past their reproductive ages leads to under-estimation of $V_{ka}/\bar{k}_a$, especially if there is a certain amount of mortality during the reproductive age group. In addition, the choice of probabilty of survival (P_s) to adulthood only increases the variance

effective size if the data are obtained from post-reproductive age men and women. Howell (1979) has given the data on eventual reproductive success of all men and women reaching adulthood among !Kung. The data yield an estimate of $N_e v/N$ as 0.814 (table 3.17). Note that the controversy regarding P_s is also resolved in this approach. I shall use this value for estimating $\bar{t}_o$ in a later section.

3.2.4.2.2. Khoikhoi.

Khoikhoi, or Nama, as they call themselves, are a small population limited in distribution at present to Namibia and the deserts just south. They are a people displaced north from the Cape of Good Hope by the advancing tide of Dutch colonists and pushed west by the hostile Bantu-speaking groups in the east. Unlike Kalahari San the present-day Khoikhoi are a mixture of 'Oorlam' clan refugees from South Africa and Nama, the original Khoikhoi of Namibia.

Numerically and culturally, Khoikhoi are a mere shadow of their former selves (Bannister and Johnson, 1979). According to O'Callaghan (1977) their number in Namibia in 1974 was 37,000. Some earlier estimates give the size of Namas ranging from 34,806 in 1959-60 (Wellington, 1967) to 39,400 in 1966 (Department of Foreign Affairs, RSA, 1967). No data on the demographic profile of Khoikhoi are available and we shall use the

TABLE 3.18 MEAN AND VARIANCE OF PROGENY SIZE AMONG !KUNG MEN AND WOMEN AND ESTIMATION OF $N_e v/N$

		AT AGE 50			AT AGE 15	
		Dobe!Kung		All San	Dobe!Kung	
		Men	Women	Women	Men	Women
1.	Mean ($\bar{k}_b$)	5.260	4.520	3.660	4.000	3.860
2.	Variance ($\bar{v}_{kb}$)	8.930	4.380	5.290	10.360	5.140
3.	Probability of survival to adulthood (P_s)	0.510	0.530	0.630	0.510	0.530
4.	$N_e v/N$	0.849	1.008	0.836	0.710	0.919
	Average ($N_e v/N$)		0.898		0.814	

parameters estimated for the San populations for Khoikhoi also. For $\hat{\lambda}$ = 0.458, the actual population size of Khoikhoi was 15,941 individuals in 1959-60.

3.2.4.2.3. Dama.

Dama are a Nama-speaking Negroid people living as a reproductive isolate in the north-west of Namibia. Although living in the same areas as Khoikhoi and San they seem to have received little contribution from either of them (Nurse *et al.* 1976). Some of their morphological features (Knussmann, 1969; Knussmann and Knussmann, 1969-70) as also genetic features (Nurse *et al.* 1976) distinguish Dama from other Negroid populations of Namibia, Angola and Botswana, which may be attributed partly to their reproductive isolation (Nurse *et al.* 1976).

In 1960 and 1966 Dama numbered 44,353 and 50,200 people respectively (Department of Foreign Affairs, Republic of South Africa, 1967). O'Callaghan's estimates of 1970 and 1974 show respectively 66,291 and 75,000 living Dama. Their equilibrium numbers are difficult to enumerate at present although, like the Nama, Dama are now also pastoralists, their life-style resembled San until quite recently. In the absence of any other authentic data I shall use the demographic features of San for the Dama. The actual population size of Dama is thus obtained to be 0.458 2x 44,353 = 20,323 individuals in 1960.

3.2.4.2.4. Black Basarwa.

In addition to Dama two other Khoisan-speaking hunter gatherer Negroid groups from southern Africa namely Kwengo from western Caprivizipfel region of Namibia and Danisan from the northwestern area of Botswana, have been included in this study. These are also described popularly as Black Basarwa (Lee, 1979) or Black Bushmen (Gusinde, 1966; Almeida, 1965) because of their morphological features. Lee (1979) includes these groups, because of their cultural and linguistic similarities, under San although Nurse and Jenkins (1977a,b) find their genetic profile rather as non-Khoisanoid. Their population numbers are given by Lee (1979) as 2,000 and 8,000 respectively (table 3.17). For $\hat{\lambda} = 0.458$, the estimates of N for Kwengo and Danison are 916 and 3,664 respectively.

3.2.4.2.5. Sandawe.

The Sandawe of central Tanzania are a small Khoisan-speaking tribe who live in comparative isolation although this isolation is by no means absolute either geographically or culturally. At the genetic level the permeability of Sandawe boundary is more or less one-sided i.e. emigration of women to other tribes, especially Turu (Neyman, 1978) which too is a recent phenomenon (Neyman 1970, 1978; Raa, 1970). Although reported to be morphologically close to Khoikhoi (Trevor, 1947)

and having migrated from southern Africa (Coon, 1965), Sandawe are placed more close to Dama, Bantu and Yoruba, than San and Khokhoi genetically (Godber *et al.* 1976; Nei and Roychoudhury, 1981).

According to the 1957 census there were about 20,031 Sandawe within their tribal boundaries. The grand total including emigrants to distant towns was 28,309 (East Africa Statistical Department 1958). Trevor (1947) gave an approximate census of 21,000 Sandawe in 1944. Taking this earlier estimate as the mean census size, we get, for $\hat{\lambda}$= 0.458, the actual population size for Sandawe as 9,618.

3.2.4.2.6. Kgalgadi.

The Kgalgadi speak a Tsawana-related language and are found mostly along the fringes of the Kalahari, into which they are said to have retreated following the arrival of a second Negro immigrant wave represented by the ancestors of modern Rolong and Tlhaping. Their way of life has, until recently, greatly resembled that of the San (Schapera, 1953). Their numbers are difficult to estimate, though they probably run into several thousands (Nurse and Jenkins, 1977a).

3.2.4.2.7. Pygmies.

Pygmies are likely to have been a much larger group at one time, inhabiting perhaps most of

central Africa between the 12°E and 16°E longitude. At present there exist four major groups of Pygmies (Murdock, 1959):

1. Western, numbering about 27,000 living in the region between Cameroon, Congo Brazza, the Central African Republic and Gabon,
2. Central Twa, numbering about 100,000 and being considerably acculturated and mixed in the central part of Congo Kinhasa,
3. Gebera, numbering about 9,000 living in Ruanda, and Urundi and having adapted a sedentary life in the plains near Lake Kivu, and
4. Mbutis, numbering about 32,000 (later estimates are somewhat higher) located in the Ituri forest in the northeast of Congo Kinhasa, who reveal the least Negro influence in physique and culture.

In addition, there are many splinter groups in central and southern Africa (Cavalli-Sforza, 1972). I have included in the present study two of the least acculturated Pygmy groups namely Western and Mbutis, although genetic data are now available on central Twa also (see Beretta *et al.* 1977).

Population density in Pygmies is of the order of 0.2 inhabitants per km^2. Birth and death rates are not known exactly but there is some indication that infant mortality may be somewhat lower and adult mortality higher than among the neighbouring

farmers (Cavalli-Sforza 1972). No results on age structure or matrimonial migration are available so far, although graphical representation of these important demographic parameters are given by Cavalli-Sforza and Bodmer (1971) and Cavalli-Sforza (1972).

From the description of marital distances among Pygmies given by Cavalli-Sforza (1972) and estimates of Wahlundian variance (Cavalli-Sforza, 1969), it is obvious that Pygmies, like San, travel farther to find spouses (a trait partly attributed to their nomadic life-style). However, the age structures, as interpreted from the graph of age blocks given by Cavalli-Sforza (1972) do not match with those of San given by Howell (1979). About 37.87% of the population of a Pygmy group seems to be in the age group 15-44 years, compared to 45.82% of San. In the absence of any other evidence, I shall use the demographic parameters derived from San data for estimation of mutation rates in Pygmies also.

The actual population size (N) for Western and Mbuti Pygmies becomes 12,371 and 14,662 respectively. The size of the breeding group for a density of 0.2 persons per km^2, is 4,734 individuals. No data are available on the average size of an endogamous group in Pygmies, although this number certainly is much smaller than the

"magic" number of 500 (Lee and de Vore 1968).

3.2.4.3. The estimation of $\bar{t}_o$, b and c.

Two estimates of $\bar{t}_o$, the mean number of generations a mutant survives prior to extinction, have been obtained. For the 80% endogamous panmictic unit, the value of $\bar{t}_o$ is given as

$$\bar{t}_o = 2\left(\frac{N_e v}{N}\right) \log_e(2N)$$

$$= 2 \times 0.814 \times \log (2 \times 132)$$

$$= 9.078 \text{ generations}$$

In low density populations such as the African hunter-gatherers under consideration here, the probability that a rare variant will pass into neighbouring populations before becoming extinct is low. For this reason the estimate of $\bar{t}_o$ based on the actual size of the 80% endogamous unit seems to be the appropriate one to use for the estimation of mutation rate by the method of Kimura and Ohta (1969).

The two parameters of geometric offspring distribution used in estimating the probability of the drift of higher frequency alleles to singletons (P_{j1}), namely b and c, were obtained by using the data on family size distributions on !Kung San given by Howell (1979). From the mean and variance of the expected progeny size for all the men and women reaching adulthood the estimates $\hat{b}$ and $\hat{c}$, after correcting for pre-adult deaths are

calculated to be 0.2388 and 0.4365 respectively.

3.2.4.4. The laboratory data.

Much of the data used here has been collated from published genetic surveys. In addition, unpublished data on some hunter-gatherer groups from Namibia and Botswana, kindly provided by Professor T. Jenkins, is also incorporated. Table 3.19 shows the summary statistics of the data used.

The distribution of private/rare variants in different Khoisan groups is given in table 3.20. For 22 protein loci, 11 variants were designated to be private. While there is no doubt, at present, regarding the Khoisan origin of the variants like PGD^{H}, $Hb\beta^{D\ Bushman}$, $Hb\alpha^{Var\ 1}$ and $Hb\alpha^{Var\ 2}$, the inclusion of the other seven variants needs further justification because of their occurrence in non-Khoisan groups.

Two alleles of locus PGM_1, namely PGM_1^6 and PGM_1^7 occur with non-polymorphic frequencies in many African populations (Nurse *et al.* 1974; Beaumont *et al.* 1979; Godber *et al.* 1976; Hitzeroth *et al.* 1979; Santachiara-Benerecetti *et al.* 1980). For PGM_1^6 the relatively high frequency of 2.34% in Khoikhoi seems to justify its inclusion here as a private polymorphism. Two individuals, one from Hei//om and the other from Zhũ/twasi are found to possess the even

TABLE 3.19 SUMMARY OF VARIOUS STATISTICS USED FOR ESTIMATING MUTATION RATE IN AFRICAN POPULATIONS

	Population/ Group	No. of loci exam-ined	Single locus deter-minations	Actual popu-lation size (N)	Variants/locus $\hat{k}_t$	$\hat{k}_r$	$\hat{K}$	Source (see Notes)
A.	San							
1.	Hei//om	15	1,123	916	0.2000	0.1429	0.5558	1,9,10
2.	Nharo	12	1,070	3,162	-	-	-	9,10
3.	G/wi&G//ana	12	1,602	1,375	0.0833	-	0.2098	8,10
4.	!Xõ & ǂhuã:	16	1,503	916	0.0625	-	0.1527	10
5.	G!aokx'ate	22	717	~50	-	**	-	9,10
6.	G!ag!ai	15	525	2,978	-	**	-	9,10
7.	//au//en	11	1,785	1,375	0.0909	-	0.2062	9,10
8.	Zhũ/twasi	16	9,643	2,749	0.0625	0.0625	0.1059	1,5,6,7,9,10
9.	Miscellaneous	8	606	100	-	-	-	6,7,10
	Total San	22	18,574	13,517	0.4545	0.3529	1.4926	1,5-10
B.	Khoihoi	16	3,205	15,941	0.0625	-	0.7397	5,6,7,10
C.	Total Khoisan	22	21,779	29,458	0.5000	0.4706	2.5681	1,5-10
D.	Negroid hunter-gatherers							
1.	Dama	17	2,199	20,323	0.1176	-	2.6230	10,12
2.	Kwengo	14	480	916	-	**	-	11
3.	Danisan	21	1,610	3,664	0.0952	--	0.7187	10
4.	Kgalgadi	16	1,831	6,154+	0.0625	-	0.5222	15
5.	Sandawe	13	2,646	9,618	0.1538	0.1538	1.1499	4
E.	Central African Pygmies							
1.	Western	23	18,627	12,371	0.3043	0.2727	0.9691	2,3,13,14
2.	Mbuti	13	2,100	14,662	0.0769	0.0769	1.0277	2,3,13,14
	Total Pygmies	23	20,727	27,033	0.3913	0.3182	2.0177	2,3,13,14

NOTES: 1. Botha *et al*, 1972; 2. Cavalli-Sforza, 1972; 3. Cavalli-Sforza *et al*, 1969; 4. Godber *et al*, 1976; 5. Jenkins, 1972; 6. Jenkins *et al*, 1968; 7. Jenkins *et al*, 1971; 8. Jenkins *et al*, 1975; 9. Nurse *et al*, 1977; 10. Nurse and Jenkins, 1977a; 11. Nurse and Jenkins, 1977b; 12. Nurse *et al*, 1976; 13. Santachiachra-Benerecetti *et al*, 1980; 14. Vergnes *et al*, 1979; 15. Unpublished data.

* Estimates obtained after excluding loci with n<50.

** No estimate possible, for n<50 for all loci.

TABLE 3.20 NUMBER OF COPIES AND FREQUENCY OF PRIVATE AND RARE VARIANTS IN KHOISAN POPULATIONS

Variant	San Populations	Copies (% frequency)		
		San	Khoikhoi	Khoisan
$Hb\alpha^{Var\ 1}$	G/wi and G//ana (38;21.88%)*	38(1.13)*	-	38(0.98)*
$Hb\alpha^{Var\ 2}$	Zhũ/twasi (1;0.05%)*	1(0.03)*	-	1(0.02)*
$Hb\beta^{D\ Bushman}$	//au//en (16;5.99%)* Zhũ/twasi (4;0.21%), Nharo (1; 0.34%)	21(0.62)*	-	21(0.54)*
$PepA^{2}$	Present in all populations except G!ag!ai & G!aokx'ate	18(1.33)	7(1.38)	25(1.10)
$PepD^{2}$	G/wi and G//ana (3;1.16%)*	3(0.23)*	-	3(0.18)*
$PepD^{3}$	Hei//om (2;1.48)*; Zhũ/twasi (3;0.48)	5(0.38)*	-	5(0.30)*
PGD^{H}	Hei//om (1;0.81%)*	1(0.04)*	-	1(0.04)*
PGD^{R}	-	-	1(0.31)	1(0.04)
ACP_1^{C}	G/wi & G//ana (2;0.70)	2(0.14)*	1(0.33)	3(0.18)
ACP_1^{R}	Present in all populations tested	353(24.96)	65(21.67)	418(24.39)*
AK_1^{2}	Present in all groups except G!aokx'ate	93(5.26)	17(5.31)	110(5.27)
ADA_1^{2}	Zhu/twasi (1;0.16)	1(0.07)	1(0.16)	2(0.10)
PGM_1^{6}	G!ag!ai (1;1.39%); Zhũ/twasi (1;0.08%)	2(0.08)	9(2.34)*	11(0.03)*
PGM_1^{7}	Hei//om (1;0.55%)*; Zhũ/twasi (1;0.08%)	2(0.08)*	-	2(0.07)*
PGM_2^{2}	Present in all groupsexcept G!ag!ai	68(2.76)*	1(0.26)	69(2.43)*
TfD^{1}	Present in all populations tested except G!aokx'ate	468(12.29)*	4(0.79)	472(10.94)*

rarer variant allele PGM_1^7, the introduction of which from outside is unlikely.

Two variants of peptidase D, namely $PepD^2$ and $PepD^3$, are considered to be of Negroid origin (Nurse and Jenkins, 1977a). In fact, the first report on peptidase D by Lewis and Harris (1969) gave these variants respective gene frequencies of 0.0240 and 0.0205 in Negro subjects of W. Indies extraction. However, in most of the recent reports on peptidase D in Negroids, evidence for the presence of these variants is lacking. Out of 7 South African Negroid populations tested for peptidase D, only Mseleni (Zulu/Tonga) showed a polymorphic frequency of $PepD^2$ (Nurse and Jenkins, 1974). For the other 747 subjects of Bantu and Khoisan-speaking Negroids, namely Kavango (Nurse and Jenkins, 1977b), Dama (Nurse *et al.* 1976), Danisan (Nurse and Jenkins, 1977a), Riemvaasmaak (Nurse and Jenkins, 1978) and Njinga (Nurse *et al.* 1979), tested for this enzyme, no copy of the variants $PepD^2$ and $PepD^3$ was recovered. The probability of such an event is significantly low ($P=1.73\times10^{-16}$).

The only other non-Khoisan population with $PepD^2$ variants is the Griqua (Nurse and Jenkins, 1975), a mixed population with large Khoikhoi/Caucasoid admixture. On the other hand, some of the San populations reveal the presence of these

variants in appreciable numbers. On the basis of gene frequencies, $PepD^2$ is assigned to Hei//om and $PepD^3$ to G/wi and G//ana.

ACP_1^R and TfD^1 are the two other alleles reportted to be present in many African populations. In Pygmies, the frequency of ACP_1^R is about 14% (see for example Santachiara-Benerecetti *et al.* 1977, 1980; Beretta *et al.* 1977 and Vergnes *et al.* 1979). Its frequency in some western African populations has also been reported to be high by Vergnes and Gourdin (1974). Since there are two different loci of high ACP_1^R frequency in Africa, one in Khoisans and one in Pygmies, two separate mutations cannot be ruled out. However, it is in Khoisans that the frequency of this allele reaches as high as 51% in some groups with an average of 24.39% (table 3.20). On the basis of comparatively high frequency of this allele in Khoisan groups I have, therefore, assigned this allele as private to Khoisans.

The transferrin D variant attains high frequencies in Northern San of about 10-20%. In the other southern African Bantu-speaking Negroid populations its frequency is much lower (about 3-4%; see for example McDermid and Vos, 1971; Hitzeroth and Hummel, 1978) and is very low in Khoisan-speaking Negroids. For this reason TfD^1 has been assigned also to Khoisans.

The designation of the PGM_2^2 allele as private

to Khoisan may be considered rather arbitrary, since it has also been detected in appreciable numbers in various other non-Khoisan populations. For example, Danisan and Kgalgadi, who otherwise show little Khoisan admixture, show this allele in very high frequencies (9.74% and 6.85%, respectively). Interestingly, however, this variant is absent from the other non-Khoisan and hunter-gatherers of southern Africa, specially Dama and Kwengo.

The private alleles for the non-Khoisanoid, morphologically Negroid, hunter-gatherers of southern and eastern Africa are shown in table 3.21. While the presence of PGD^R in the other south African populations indicates Dama admixture, the designation of AK_1^2 and ACP_1^C as Dama and Kgalgadi variants respectively, on the basis of relative frequencies and assumption of parallel mutations, needs further clarification.

Frequencies of AK_1^2, as high as those of Dama (8.82%), //au//en (8.54%) are unusual outside the Indian subcontinent (Tills *et al.* 1971). Considering that the Dama received little or no genetic contribution from Khoisans (Nurse *et al.* 1976) and that they have been absorbed by Khoikhoi and San in large numbers, the allele is consider to be of Dama origin (Nurse and Jenkins, 1977a).

On the other hand, ACP_1^C, an allele quite common in Caucasoids is detected in very low

TABLE 3.21 NUMBER OF COPIES AND FREQUENCY OF PRIVATE AND RARE VARIANTS IN NON-KHOISAN NEGROIDS

Variant	COPIES (% FREQUENCY)				
	Dama	Kwengo	Danisan	Kgalgadi	Sandawe
$Hb\alpha^{B}2$	-	-	-	10(4.03)	-
GPX^{Fast}	**	**	11(14.29)*	**	**
LDH^{Fast}	**	**	**	**	3(0.72)*
$PepA^{2}$	4(2.17)	8(11.43)	7(4.55)	10(4.03)	**
ICD^{4}	**	**	2(1.30)*	**	**
PGD^{R}	7(2.43)*	-	-	1(0.40)	2(0.47)
PHI^{Var}	**	**	**	**	7(1.65)
ACP_{1}^{C}	1(0.54)	-	-	4(1.61)*	-
ACP_{1}^{R}	1(0.54)	6(8.57)	8(5.19)	26(10.48)	-
AK_{1}^{2}	21(8.82)*	**	-	2(0.81)	3(0.74)
CA_{2}^{2}	**	**	9(5.84)	9(3.63)	**
ADA^{2}	1(0.26)	-	-	**	**
PGM_{1}^{7}	-	-	-	-	2(0.49)*
PGM_{2}^{2}	-	-	15(9.74)	17(6.85)	3(0.74)
TfD^{1}	1(0.35)	1(1.43)	4(2.53)	6(5.56)	20(4.72)

* Private.

** Not tested.

numbers in populations of southern Africa. In Kgalgadi, however, the allele reaches polymorphic frequencies. It is unlikely that Kgalgadi, with no evidence of other Caucasian markers will have derived this character in such high frequency by admixture. Similar arguments for a new focus of ACP_1^C in the Negro populations of southern Angola and northern Namibia are given by Nurse and Jenkins (1977a).

The two private variants of Sandawe, namely PGM_1^7 and *LDH* ^{Fast} have been assigned similarly on the basis of isolated recovery of variants. The polymorphic variant of PHI has, however, been excluded because of its presence in high frequencies in the neighbouring Nyaturu, a Bantu-speaking Negroid group of central Tanzania (Godber *et al.* 1976).

For 23 loci, as many as 9 private variants have been recovered in the Pygmy populations (table 3.22). Two of these, namely PGM_2^{6Pyg} and $Hb\alpha^{Flatbush}$ are, in fact, present in polymorphic frequencies.

The Ibadan variant of G-6-PD has been reported in a number of other populations from western Africa and the Sahara (Vergnes *et al.* 1978). However, although Gd^{Ibadan} has only low frequencies in Pygmy populations, because outside groups are thought not to have contributed to the

TABLE 3.22 NUMBER OF COPIES AND FREQUENCY OF PRIVATE AND RARE VARIANTS IN PYGMIES FROM CENTRAL AFRICA

Variant	Western Pygmies		Mbuti		All Pygmies	
	Copies	Freq.(%)	Copies	Freq.(%)	Copies	Freq.(%)
$Hb\alpha^{Flatbush}$	47	1.87*	-	-	47**	1.69*
$Hb\alpha^{Babinga}$	15	0.60*	-	-	15**	0.54*
$PepA^{3}$	-	-	2	0.49*	2	0.14*
Gd^{Ibadan}	7	0.48*	-	-	7	0.48*
PGM_{1}^{4}	2	0.06*	-	-	2	0.05*
PGM_{2}^{4Pyg}	10	0.31*	-	-	10	0.28*
PGM_{2}^{6Pyg}	173	5.30	26	7.43	199	5.51*
PGM_{2}^{9}	3	0.09*	-	-	3	0.08*
$PepC^{2}$	9	0.28*	-	-	9	0.71*

* Private

** Estimated from Cavalli-Sforza's (1972) data.

Pygmy gene pool, it is considered that this allele is a Pygmy mutation.

3.2.4.5. Estimation of k_t, k_r and K.

Table 3.19 shows the estimates of the number of alleles recovered per locus in samples ($\hat{k}_t$ and $\hat{k}_r$) and the estimated values, extrapolated to the size of the total population ($\hat{K}$). The estimates $\hat{K}$ have been obtained by using the extrapolation equation of Rothman and Adams (1978). The values of $\tilde{g}(j)$, the relative proportions of various rare alleles by the number of copies were obtained from the observed distributed for the total Khoisan rare alleles. The vector of g(j) so obtained is, however, significantly different from such other estimates obtained earlier on the populations from Australia, Papua New Guinea and India.

3.2.4.6. Estimates of mutation rate

The estimates of mutation rate given by the indirect methods, $\hat{\mu}_{K-O}$ (Kimura and Ohta, 1969), $\hat{\mu}_{NEI}$ (Nei, 1977) and μ_{R-A} (Rothman and Adams, 1978) for various hunter-gatherer populations of southern and central Africa are given in table 3.23. Because no private alleles were recovered for G!aokx'ate, G!ag!ai and Kwengo, possibly because of small sample sizes, no estimates of μ were obtained. Nei's method is not applicable in a further seven populations because either sample sizes were less than 1/q or of non-recovery

TABLE 3.23 ESTIMATES OF MUTATION RATE ($\times 10^6$) IN VARIOUS AFRICAN POPULATIONS

	Population/ Group	$\hat{\mu}_{K-O}$	$\hat{\mu}_{NEI}$	$\hat{\mu}_{R-A}$
A.	San			
1.	Hei//om	33.42	82.51	10.22
2.	Nharo	-	-	-
3.	G/wi & G//ana	8.40	-	2.57
4.	!Xõ & ≠huã:	9.18	-	2.81
5.	G!aokx'ate	-	**	-
6.	G!ag!ai	-	**	-
7.	//au//en	8.26	-	2.53
8.	Zhũ/twasi	2.12	2.28	0.65
	Average	7.67	14.13	2.35
B.	Khoisans			
1.	Total San	6.08	2.12	1.86
2.	Total Khoikhoi	2.56	-	0.78
3.	Total Khoisan	4.80	1.23	1.47
C.	Negroid hunters-gatherers			
1.	Dama	7.11	-	2.17
2.	Kwengo	-	**	-
3.	Danisan	10.80	-	3.31
4.	Kgalgadi	4.67	-	1.43
5.	Sandawe	6.58	2.85	2.01
	Average	5.83	0.71	1.78
D.	Central African Pygmies			
1.	Western	4.32	1.98	1.32
2.	Mbuti	3.86	1.12	1.18
	Total Pygmies	4.11	1.02	1.26

** No result.

of rare alleles for some of these populations which affect the average estimates of $\hat{\mu}_{NEI}$ considerably.

The estimates of μ obtained by the three methods for total Khoisans, non-Khoisanoid morphologically Negroid hunter-gatherers and total Pygmies are quite similar. The average over the three methods for each population yields values of mutation rate of 2.50×10^{-6}, 2.77×10^{-6} and 2.13×10^{-6} /locus/generation respectively.

The estimates of mutation rate obtained for individual groups are on average higher than those obtained by pooling the data over total populations. The average estimates for individual values are inflated by some particularly high estimates of μ, as in the Hei//om. It may, however, be noted that the average value obtained by pooling the data is relatively more efficient.

3.2.4.7. Discussion.

The estimates of mutation rate generated by the three procedures used here in hunter-gatherers of Africa are intermediate between such estimates generated for the south Indian Scheduled Tribes (table 3.16) and the Papua New Guinean tribes (table 3.9). However, the estimates generated by the method of Rothman and Adams (1978) in these populations are consistently lower than those obtained by this method for other world populations.

This bias toward lower values in the mutation rate estimates is mainly because of the lower value of $\tilde{g}(1)-\Sigma\tilde{g}(j)P_{j1}$, obtained for the African populations (0.0337). Similar estimates of $\tilde{g}(1)-\Sigma\tilde{g}(j)P_{j1}$ obtained for Australian Aborigines, Papua New Guineans and South Indian Scheduled Tribes are 0.2177, 0.2770 and 0.2479 respectively. Although Rothman and Adams (1978) recommend the use of a vector of $\tilde{g}(j)$ estimated from the demographic data, the above estimates have been obtained from the observed distribution of rare allales recovered in large samples over sets of loci. Except for the estimates of African populations, all the other estimates obtained earlier are quite close to the estimate of 0.2096 generated from demographic data on Yanomama by Neel and Rothman (1978).

However, it is acknowledged that the relative proportions of P_{ji} generated by a set of b and c values represent the realtive distribution of alleles only in the present generation. If the population is undergoing any demographic changes, this distribution becomes redundant (Rothman and Adams, 1978; Neel *et al.* 1980c). Similar trends will be reflected in the estimates of g(j). However, despite these expected variations, the low value of 0.0377 in African populations is an aberrant phenomenon.

The extraordinarily high estimates of μ obtained for the Hei//om are attributable to their overlap in distribution with the Zhũ/twasi group. Nurse and Jenkins (1977a) have also referred to the possible gene flow between Hei//om and Zhũ/twasi. Since two of the variants considered private to Hei//om ($PepD^3$ and PGM_1^7) are also present in the much larger Zhũ/twasi populations (table 3.20), the estimates of μ in Hei//om may be considered provisional.

Standard errors for these estimates are not given. This is because the variation in the number of alleles among loci and also the errors associated with estimating the actual size of populations has to be taken into consideration. In addition, in Kimura and Ohta's (1969) method the variance due to the estimate of $\bar{t}_o$ is quite large.

The estimates of mutation rate obtained on African populations may be viewed with certain cautions. One of the most probable sources of error is the assumption of neutrality. Appreciable amounts of selection could bias the estimates in either direction. However, the choice of rare alleles in the method of Nei and singletons in that of Rothman and Adams ensures that the role of selection in maintaining the alleles biases these estimates to a minimum.

The rates of mutation obtained by disregarding the structure of population are on an average lower than those obtained from individual groups. For neutral genes such pooling should have led to no differences in the estimates (Nei, 1977). Such differences, however, are expected if the deleterious genes exist among the set of loci. However, it would be premature to generalize from these differences in African populations to all human populations.

3.2.5. Some additional estimates.

The indirect estimates of mutation rate in the populations from Australia, Papua New Guinea and India, reported by Bhatia *et al*. (1979), Bhatia *et al*. (1981) and Bhatia (1981b) respectively, were obtained by the then available procedures of Kimura and Ohta (1969), Nei (1977) and Rothman and Adams (1978). These results have been presented, with slight modifications, in the original format in sections 3.2.1. to 3.2.3. To maintain uniformity in the presentation, the results on African populations have also been obtained by the same procedures and presented in section 3.2.4. in almost similar style.

Recently, another procedure for indirect estimation has been suggested by Chakraborty (1981) on the basis of the number of rare alleles recovered in a sample ($\hat{k}_r$). Extending this approach to singletons recovered in a sample ($\hat{k}_s$) a new method of moment estimator has been suggested in Chapter 2. To update my previous results I have estimated the rate of mutation by these two new methods for all the populations reported in the previous sections. To keep the section self-sufficient, the results are presented in the form of a report.

3.2.5.1. Introduction,

The estimation procedures of Chakraborty (1981) and the singletons approach are summarized as follows:

1. Chakraborty's method. The estimator is given as

$$\hat{\mu}_{CHAK} = \frac{A-\sqrt{A^2-4Bk_r}}{8NB} \qquad (3.7)$$

where k_r is the number of alleles recovered in a sample of 2n genes and N is the actual population size. A and B are given as

$$A = \sum_{j=1}^{[2nq]} j^{-1} \qquad \text{and}$$

$$B = \sum_{j=1}^{[2nq]} (2n-j)^{-1}$$

where [2nq] is the largest integer value equal to or less than the inner expression. For large [2nq], A is approximated to log [2nq]+ γ, where γ is Euler's constant and B is approximated to -log [1-q].

2. Singletons method. This method estimates the mutation rate as

$$\hat{\mu}_{k_s} = \frac{k_s(2n-1)}{4N(2n-k_s)} \qquad (3.8)$$

where k_s is the number of singletons recovered in the sample of 2n genes.

3.2.5.2. Results.

The summary statistics of the data required in obtaining the estimates for various populations are given in table 3.24. Only populations with non-null results for k_r or k_s have been included.

Table 3.25 shows the estimates of mutation rate obtained by the two methods. In addition, the estimates of mutation rate obtained earlier by Nei's method have also been listed for comparison. In Toda and Hei//om the estimates of μ by Nei's methods are more than two fold those obtained by the approach of Chakraborty. This is because Nei's method invariably leads to over-estimation of mutation rate when sample sizes are small (Chakraborty, 1981), especially when $2nq<2.7183$. As noticed by Chakraborty the methods of Chakraborty and Nei lead to similar results for large sample sizes, although $\hat{\mu}_{CHAK}$ is always smaller than $\hat{\mu}_{NEI}$.

The singletons method in comparison leads to fewer estimates because of the failure to recover singletons in some population samples. The estimates of mutation rate obtained by the singletons method are still on an average higher, despite the fewer observations, than those given by Chakraborty's method.

The probability of non-recovery of any

TABLE 3.24 SUMMARY STATISTICS OF THE DATA USED FOR ESTIMATING THE MUTATION RATE FOR VARIOUS POPULATIONS

Population/group	No. of peptides examined	Mean No. of genes sampled ($2\bar{n}$)	Actual population size (N)	Variants/locus $\hat{k}_r^+$	Variants/locus $\hat{k}_s^+$
A. Individual populations*					
1. Karkar Island (PNG)	18	1,822	3,735	0.111	0.111
2. Siassi Island (PNG)	17	538	2,310	0.118	0.118
3. Koya (SI)	19	831	13,297	0.211	0.105
4. Rajgond (SI)	17	420	8,136	0.059	-
5. Gadaba (SI)	16	1,629	1,264	0.063	-
6. Yerukula (SI)	17	67	4,357	**	0.059
7. Toda (SI)	22	209	416	0.045	0.045
8. Kadar (SI)	22	425	681	0.045	0.045
9. Hei//om (NAM)	15	150	916	0.143	0.133
10. Zhũ/twasi(BCT)	16	1,205	2,749	0.063	0.063
11. Sandawe (TAN)	13	407	9,618	0.154	-
12. W. Pygmies (C.AFR)	23	1,620	12,371	0.273	-
13. Mbuti (C.AFR)	13	323	14,662	0.077	-
B. Regional populations					
1. N. Andhra tribes (SI)	29	2,632	52,560	0.241	0.069
2. S. Andhra tribes (SI)	22	524	12,085	0.045	0.045
3. Kerala/TN tribes (SI)	27	1,769	52,002	0.185	0.074
4. San (NAM, BOT)	22	1,768	13,517	0.353	0.091
C. Continental populations					
1. Australian Aborigines	25	5,214	9,160	0.520	0.240
2. Papua New Guineans	21	12,026	34,450	0.952	0.524
3. S. Indian Sch. Tribes	30	4,521	116,089	0.467	0.167
4. Khoisans	22	1,980	29,458	0.471	0.091
5. Pygmies	23	1,802	27,033	0.318	-

Note: * 25 additional populations studied have not been included because of the recovery of no rare/private alleles.

** sample size <50

+ The estimates of k_r based only on loci with n>50.

TABLE 3.25 ADDITIONAL RESULTS OF μ OBTAINED BY USING CHAKRABORTY AND SINGLETONS METHOD

Population/groups*	$\mu \times 10^6$		
	$\hat{\mu}_{NEI}$	$\hat{\mu}_{CHAK}$	$\hat{\mu}_{k_s}$
A. Individual populations[a]			
1. Karkar Island	2.71	2.64	7.44
2. Siassi Island	7.56	5.36	12.70
3. Koya	1.87	1.53	1.98
4. Raj Gond	1.26	0.94	-
5. Gadaba	4.43	3.71	-
6. Yerukula	**	**	1.09
7. Toda	37.00	18.60	27.07
8. Kadar	11.55	8.43	16.61
9. Hei//om	82.51	39.05	36.09
10. Zhũ/twasi	2.28	1.87	5.72
11. Sandawe	2.85	1.92	-
12. W. Pygmies	1.98	1.65	-
13. Mbuti	1.12	0.72	-
Average	4.62[b]	2.54[b]	2.86[c]
B. Regional populations			
1. N. Andhra tribes	0.35	0.30	0.33
2. S. Andhra tribes	0.57	0.49	0.94
3. Kerala/TN tribes	0.31	0.25	0.36
4. San	2.12	1.92	1.68
Average	0.84	0.74	0.83
C. Continental populations			
1. Australian Aborigines	3.58	3.08	6.51
2. Papua New Guineans	1.44	1.30	3.80
3. S. Indian Scheduled Tribes	0.26	0.23	0.36
4. Khoisan	1.23	1.14	0.77
5. Pygmies	1.02	0.85	-
Average	1.51	1.32	2.29

* Only populations with non-null results of $\hat{k}_r$ and $\hat{k}_s$ included.

** No result for μ_{NEI} and μ_{CHAK} possible because of $2n<100^{NEI}$.

[a] 25 additional populations with no result for k_r or k_s.

[b] averaged over 34 populations. The methods of Chakraborty and Nei are inapplicable in 4 of the 38 of these populations because of $n<50$.

[c] averaged over 38 populations.

variant in a sample decreases with less conservative upper limits (q). The lack of singletons in Raj Gond, Gadaba, Sandawe, W. Pygmies, Mbuti, as also in the total sample of Pygmies, may be viewed in this regard. The rare allele approach is, however, available only with samples of sizes more than 1/q genes. No estimate of mutation rate was thus possible in Yerukula ($2\bar{n}$=67), although a private variant, PGM_2^{10}, was recovered as a single copy (Bhatia, 1981; Blake *et al.* 1981) which yields a value of $\hat{\mu}_{k_s}$ as 1.09×10^{-6} per locus/generation.

The inconsistencies in the estimates of μ for individual populations, therefore, are largely an artefact of small sample sizes. The results, however, exhibit less variability when the populations are pooled on a regional or continental basis and the sample size is sufficient. The mean estimates of mutation rate on a continental basis by the rare allele methods of Nei and Chakraborty and the singleton approach are 1.51×10^{-6}, 1.32×10^{-6} and 2.29×10^{-6} per locus/generation respectively.

3.2.5.3. Discussion.

Two important observations can be made from the results of mutation rates given in table 3.25. Firstly, the choice of singletons leads on an average to higher estimates of μ than those obtained by the use of rare alleles. It may be

that the choice of q influences the estimates of μ in some way. Secondly, it is clear that there is a paucity of singletons in African populations.

To illustrate the role of an arbitrary upper limit (q) in influencing the estimates of mutation rate, the sampling equations of Nei (1977) and Chakraborty (1981) have been employed for various subsets of allelic data. The estimates of μ in Australian Aborigines, Papua New Guineans and South Indian Scheduled Tribes (table 3.26) exhibit a decline in the value of μ with the moving upward of the limit for defining variants. The largest estimates are obtained when only singletons are used.

Nei (1977) has pointed out that in the presence of deleterious mutations, the use of a small value of q with large sample size also estimates the rate of deleterious mutations including other types of mutations. For large values of q an under-estimate of mutation rate will be obtained by his formulation if deleterious genes are present. Accordingly small values of q lead to estimates of the total mutation rate in such situations. The choice of singletons ($q = \frac{1}{2n}$) , thus measures the total rate of mutation if deleterious genes are present and the sample size is large.

One of the problems, alluded to in section 3.2.4 earlier, is the very low recovery of singletons in hunter-gatherer populations from Africa. Compared

TABLE 3.26 ESTIMATES OF MUTATION RATE ($x10^6$) FOR DIFFERENT VALUES OF q BY THE METHODS OF NEI (1977) AND CHAKRABORTY (1981)

q	Australian Aborigines		Papua New Guinea		South Indian Tribes	
	Nei	Chakraborty	Nei	Chakraborty	Nei	Chakraborty
1.000	2.04	2.02	0.81	0.80	0.12	0.11
0.050	3.14	2.90	1.13	1.04	0.19	0.17
0.010	3.59	3.22	1.44	1.34	0.26	0.23
0.005	3.68	3.24	1.52	1.39	0.32	0.28
0.001	5.95	4.82	2.08	1.86	0.33	0.26
1/2n*	5.96**	6.51	3.52	3.80	0.39**	0.36

* n is the number of individuals sampled.

** Evaluated for $\theta \int_{.5}^{1.5} x^{-1} dx$

to the values of 0.462, 0.550 and 0.358 for the ratio k_s/k_r in Australian Aborigines, Papua New Guineans and South Indian tribes respectively, the ratio in Khoisans and Pygmies is 0.193 and zero respectively. One of the reasonsfor this aberrancy is the relatively positive growth trends in Australian Aborigines, Papua New Guineans and South Indian tribes while South African populations are yet to recover from their earlier drop in numbers.

One advantage in using the singletons approach is its applicability to populations with sample sizes smaller than 100 genes for q=0.01 or smaller than 1,000 genes with q=0.001. This is so because the singletons approach is not exactly an extension of the rare allele method.

Chapter 4

FACTORS AFFECTING ESTIMATION OF ELECTROMORPH MUTATION RATES

4.1. Introduction.

In the previous chapter, the average electromorph mutation rates for different sets of loci have been obtained for various world populations. A number of factors, however, could affect these estimates of mutation rates. These include the choice of protein loci for which electrophoretic data are available, the variability of mutation rates among loci and the different aspects of allelic data which are used as input. In addition, the role of sample size could affect the estimates of mutation rates obtained by different methods.

To illustrate the role of these factors, two sets of data are analyzed in this chapter. In section 4.2. the estimates of mutation rate in Australian Aborigines are re-examined for the effects of sample size, molecular size and structure and heterozygosity. In section 4.3. the estimates of mutation rate on Amerindians obtained by Neel and Rothman (1978) and by Chakraborty (1981) are compared to determine the effects of their different sampling algorithms on the estimates.

4.2. Factors Affecting Estimation of Electromorph Mutation Rates in Australian Aborigines.

During the past decade a number of statistical methods to calculate the mutation rates at cistron level from electrophoretic data, both direct (Mukai, 1970; Mukai and Cockerham, 1977) and indirect (Kimura and Ohta, 1969; Nei, 1977; Rothman and Adams, 1978), have been developed. The latter methods are based on the detection of private electrophoretic variants in random samples from isolated populations. Some of the assumptions made are: (1) that there is a complete one-to-one correspondence between the incidence and detection of rare electromorphs, (2) that all the rare alleles observed in the population are introduced and maintained through mutation only, and (3) that there is a constancy of mutation rates, on an average, over any subset of protein and enzyme loci.

The class of relationship given by (1) is very difficult to evaluate as estimates of number of alleles depend critically upon sample size (Harris *et al.*, 1974; Koehn and Eanes, 1978; Eanes and Koehn, 1978; Bhatia *et al.*, 1979) and upon the resolution of the experimental techniques employed to discriminate allelic variants (Johnson, 1977a). The last point is not trivial as new techniques suggest that there exists a large reservoir of previously undetected alleles

(Johnson, 1977b). Introduction of new alleles by sources other than mutation, e.g., intragenic recombination, was suggested by Watt (1972), Koehn and Eanes (1976) and Strobeck and Morgan (1978).

The assumption included in (3) is the weakest since inter-locus variability in mutation rates has been noted by Nei *et al.* (1976b). On the basis of aminoacid substitutions in various polypeptide chains, they found this variability to follow the gamma distribution. Zouros (1979) pointed out that over a large range of species only certain types of enzymes occupy the same tail of the distribution, indicating the role of physicochemical features of the molecules and this may explain, in part, the inter-locus variability in mutation rates.

Parameters of genetic variation, like heterozygosity and the number of rare alleles, are affected by a number of factors. For heterozygosity these include:- substrate specificity (Gillespie and Kojima, 1968), physiological function (Johnson, 1974), quaternary structure (Zouros, 1976; Harris *et al.*, 1977; Ward, 1977) and subunit size (Koehn and Eanes, 1977; Nei *et al.*, 1978; Brown and Langley, 1979). The relationship among these has been demonstrated for both invertebrate and non-human vertebrate

species. However, Harris *et al.* (1977) have detected lack of correlation between subunit molecular weight and heterozygosity in European human populations. Nei *et al.* (1978) attributed this to the low level of mean heterozygosity in human populations.

In the case of the number of rare alleles the factors include:- effective population size (Ohta, 1972; Rothman and Adams, 1978), intragenic recombination (Morgan and Strobeck, 1979), subunit size (Eanes and Koehn, 1978), founder effect (Thompson and Neel, 1978), polymorphism (Harris, 1975), bottleneck effect (Bhatia *et al.*, 1979) and transient distribution of neutral alleles (Nei and Li, 1976). The list is by no means exhaustive and a whole set of cause-effect factors, which include the total number of alleles segregating at a locus, mean level of heterozygosity and subunit number etc. can be included for their role in the introduction and maintenance of rare alleles in a population. Since the estimation of mutation rates by indirect methods depends on the number of rare alleles, it is important to reassess the role of the above factors in determining these rates. In addition, because of the correspondence between molecular weight and mutation rates and the former's role in introducing interlocus variability in mutation rates

at the peptide level (Nei *et al.*, 1976b), it may be relevant also to calculate the mutation rates at base pair level (Mukai and Cockerham, 1977), making cistronic comparisons independent of molecular weight.

4.2.1. The data.

Most researchers who have studied the role of variability in mutation rate and heterozygosity, because of the difficulty of controlling all the factors involved, restricted themselves to answer only one or two queries. They compensated for the lack of control by increasing the range of species for which results were given. However, an ideal choice for an answer is a subdivided population, distributed over a large geographical area and sampled extensively. The electrophoretic results for Australian Aborigines reviewed by Blake (1979), seem to provide an adequate set of data for analysis. Blake's data as retabulated by Bhatia *et al.* (1979) has been used in the present study. The loci included, arranged into monomers and multimers, and their respective sample sizes, are shown in table 4.1. The multimer loci are all dimers except for the LDH loci. The subunit molecular weights have been taken from the tabulation by Hopkinson *et al.* (1976). A total of 15 multimeric and 10 monomeric loci have been included. In the absence of any direct relation-

TABLE 4.1 LIST OF PROTEINS AND ENZYMES INCLUDED IN THE STUDY AND THEIR RESPECTIVE SAMPLE-SIZES, SUBUNIT SIZES, NUMBER OF TOTAL AND RARE ALLELES AND EXPECTED HETEROZYGOSITY

Enzyme System	Abbreviation	No. of individuals sampled	Subunit* size (in daltons)	Total number of alleles	Total number of rare alleles	Heterozygosity ($1-\Sigma x_i^2$)
A. Multimerics**						
Hemoglobin- α	Hb- α	2692	15,000	1	-	-
Hemoglobin- β	Hb- β	2692	16,000	1	-	-
Superoxide dismutase	SOD_A	1795	16,000	1	-	-
Glyoxalase	GLO	1290	24,000	2	-	0.0380
Esterase D	EsD	1556	28,000	2	-	0.1467
Malate dehydrogenase	MDH	2964	35,000	1	-	-
Lactate dehydrogenase- A	LDH_A	4180	35,000	2	1	0.0002
Lactate dehydrogenase-B	LDH_B	4180	35,000	2	1	0.0004
Glutamic oxalaceticacid transaminase	GOT	748	46,000	1	-	-
Peptidase A	Pep A	3034	46,000	2	1	0.0008
Isocitrate dehydrogenase	IcD_s	1226	48,000	1	-	-
Glutamic pyruvic transaminase	GPT	1391	50,000	2	-	0.3211
6-Phosphogluconate dehydrogenase	6-PGD	4035	52,000	3	1	0.1031
Glucose-6-phosphate dehydrogenase	G-6-PD	1014	53,000	2	1	0.0010
Phosphohexose isomerase	PHI	1569	62,000	2	1	0.0006
B. Monomerics						
Acid phosphatase-1	ACP_1	4016	15,000	4	1	0.0675
Adenylate kinase-1	AK_1	3535	22,000	1	-	-
Carbonic anhydrase-1	CA_1	3751	29,000	3	2	0.0516
Carbonic anhydrase-2	CA_2	3751	29,000	2	1	0.0425
Diaphorase	DIA	1861	30,000	1	-	-
Adenosine deaminase	ADA	1437	34,000	2	-	0.0309
Phosphoglycerate kinase	PGK	1569	50,000	1	-	-
Phosphoglucomutase-1	PGM_1	3919	51,000	4	2	0.2097
Peptidase B	Pep B	3189	55,000	3	2	0.0208
Phosphoglucomutase-2	PGM_2	3790	61,000	3	2	0.0284

* After Hopkinson et al (1976)

** Except LDH loci, which are tetramers, all multimeric loci are dimers.

ship between subunit number and subunit size (Hopkinson *et al.*, 1976), the data for subunit sizes were also pooled together.

The electromorph mutation rates per cistron per generation were calculated by using the methods of Kimura and Ohta (1969) and Nei (1977). The rates at cistron level were then converted to mutation rates per base pair per generation as suggested by Mukai and Cockerham (1977) with only a slight modification. The mutation rates for multimers were computed by subtracting 14% and 28% from the total number of base pairs for dimers and tetramers respectively. This accounts for the aminoacid residues involved in surface interactions (Turner *et al.*, 1979).

The coefficients of correlation between various parameters were computed by using both the Spearman's rank order non-parameteric and Pearson's product moment correlations. Whenever required, the variables were log-transformed to equalize and normalize the variances. The analysis was performed by structuring different classes within each category to equalize or isolate the role of a particular factor.

4.2.2. Results.

Table 4.2 shows the distribution of loci at which private variants were detected. The data

TABLE 4.2 MEANS AND S.D.'S OF SAMPLE SIZE, SUBUNIT SIZE AND HETEROZYGOSITY AT LOCI WITH OR WITHOUT RARE ALLELES

	Type of Enzyme	No. of cistrons	Sample Size		Subunit Size		Heterozygosity	
			Mean	S.D.	Mean	S.D.	Mean	S.D.
Loci with rare alleles	Multimers	6	3002	1403	47,167	10,720	0.0177	0.0418
	Monomers	6	3736	288	40,000	18,188	0.0668	0.0709
	Total	12	3369	1039	43,583	11,028	0.0422	0.0612
Loci without rare alleles	Multimers	9	1817	780	30,888	14,385	0.0562	0.1104
	Monomers	4	2100	972	34,000	11,775	0.0077	0.0154
	Total	13	1904	257	31,846	13,222	0.0413	0.0934
Total	Multimers	15	2291	1188	37,533	14,875	0.0408	0.0893
	Monomers	10	3082	1036	37,600	15,479	0.0431	0.0617
	Total	25	2607	1176	37,560	14,796	0.0417	0.0780

have been classified into two categories, namely multimers and monomers to avoid the role of functional constraints in influencing other factors. Although the difference is small between multimers and monomers with respect to subunit size (mean values and S.D.'s are 37.53 $\pm$ 14.85 and 37.60 $\pm$ 15.48) and mean expected heterozygosity (0.0408 and 0.0431) respectively, the retention of these divisions is relevant for other comparisons.

4.2.2.1. Relationship between the number of rare alleles and:

(1) Sample size: Eanes and Koehn (1978) and Bhatia *et al.* (1979) showed that the efficiency of estimates of mean number of electrophoretic alleles increases with sample size. This was observed also in the present study. The product moment correlation of the total number of alleles as well as the total number of rare alleles with sample size was significantly positive ($r=0.537$, d.f. 23, $P<0.003$ and $r=0.625$, d.f. 23, $P<0.001$ respectively). The relationship showed better correspondence in monomers ($r=0.667$, d.f. 8, $P<0.018$ and $r=0.710$, d.f. 8, $P<0.011$ respectively) but the correlation with multimers was significant for rare alleles only ($r=0.329$, d.f. 13, $P<0.116$ and $r=0.506$, d.f. 13, $P<0.027$).

(2) Total number of alleles: A significant correlation exists between the total number of alleles and the number of rare alleles because one is included in the other data set. The mean value of Pearson's coefficient for this correlation was significant at 0.1 per cent level of probability ($r=0.803$, d.f. 23, $P<0.001$). But since it is an analysis of cause-effect relationship, the results can be appreciated better if some variables which affect both of them simultaneously are standardized. The partial correlations by controlling the sample size and mean amount of heterozygosity, individually and combined, yield similar high relationships, although in monomers, controlling by sample size is non-significant.

(3) Heterozygosity: Since the mean amount of heterozygosity per locus in any population is a function of the total number of alleles, a correlation between the two is to be expected. According to the stepwise mutation model and the intragenic recombination model, the introduction of new alleles will depend upon the frequencies of existing alleles, which is measured by heterozygosity. In the present data the estimates of mean heterozygosity and its variance are 0.042 and 0.006 respectively. Spearman's rank order correlations

for heterozygosity with rare alleles and total number of alleles are significant ($r=0.498$, d.f. 23, $P<0.01$ and $r=0.852$, d.f. 23, $P<0.01$ respectively). The product moment correlation between number of variants and heterozygosity shows a negative correlation (significant at 1 per cent level of probability) if the values are controlled for total number of alleles. This suggests that the number of rare alleles as a function of heterozygosity or of the total number of alleles, as inferred from the stepwise mutation model, is misleading, particularly for low values of mean heterozygosity.

(4) Subunit number: Table 4.2 shows the distribution of loci at which rare alleles were detected in terms of monomeric and multimeric loci. Whereas about 60 per cent of the monomeric loci exhibit the presence of rare alleles the fraction is 40 per cent in multimers. The number of rare alleles per locus is also much higher in monomers than in multimers (1.00 against 0.40 per locus: table 4.3). This indicates that a negative association between the number of subunits and rare alleles exists and for the log-transformed variables the present data shows a significant negative correla-

tion (r-0.483, d.f. 23, P<0.007).

(5) Subunit size: Eanes and Koehn (1978) obtained significant correlations between the subunit size and total number of alleles at enzyme loci in human populations. Since Harris *et al.* (1977) found no correlation between subunit size and heterozygosity this suggests a direct relationship between subunit size and the number of rare alleles. In the present data the correlation between the total number of alleles and subunit size is low but the number of rare alleles show a significant relationship (r=0.463, d.f. 23, P<0.01). The partial coefficient of correlation between the total number of rare alleles and subunit size is increased significantly when controlled for sample size (r=0.666, d.f. 22, P<0.001). It is clear that rare alleles are strongly correlated with subunit size when other factors are standardized.

4.2.2.2. Effect on mutation rates.

From the relationships outlined above, it is obvious that there are several factors which influence the number of rare alleles. I have, therefore, recalculated the electromorph mutation rates from the data for Aborigines following the

methods of Kimura and Ohta (1969) and Nei (1977).

Table 4.3 shows the relationship between the sample size and the estimated average mutation rates. The average number of rare alleles per locus is much higher in sample sizes above 3000 than below 3000 (1.27 against 0.14). This results in a 9-fold difference between these two sample sizes when mutation rates are calculated by the method of Kimura and Ohta (1969). For the purpose of comparison, three categories of $n>3000$, $n<3000$ and all sample sizes were made. The results show a systematic decrease in mutation rates in these respective categories.

The second important factor which operates to influence the incidence of rare alleles is the presence of polymorphism at a particular locus. The difference between mutation rates for the polymorphic and non-polymorphic loci is almost twofold indicating the fact that the stepwise mutation model can be invoked to explain these differences (table 4.4). The difference between the multimer and monomer subgroups could not be given weight because of differences of sample sizes and the incidence of heterozygosity. The results in table 4.5 show the mutation rates per cistron/generation for three different categories of subunit size, each further subdivided into

TABLE 4.3 SAMPLE SIZE AND ELECTROMORPH MUTATION RATES

Sample size	Type of enzyme	Mean sample size	Mean subunit size	No. of cistrons	Total no. of rare alleles	μ per cistron (x106)	
						Kimura and Ohta's method	Nei's method
≥3,000	Multimers	3657	42,000	4	4	16.02	6.36
	Monomers	3707	37,428	7	10	22.93	6.34
	Total	3689	39,091	11	14	20.43	6.35
<3,000	Multimers	1794	35,727	11	2	2.92	1.39
	Monomers	1623	38,000	3	-	-	-
	Total	1757	36,214	14	2	2.28	1.09
All	Multimers	2291	37,400	15	6	3.35	3.33
	Monomers	3082	37,600	10	10	19.26	4.64
	All	2607	37,480	25	16	8.67	3.60

TABLE 4.4 AMOUNT OF HETEROZYGOSITY AND ELECTROMORPH MUTATION RATES IN THE AUSTRALIAN ABORIGINES

Type of loci	Proportion of heterozygosity ($1-\Sigma x_i^2$)	Enzyme structure	Mean heterozygosity	Mean sample size	Mean subunit size (in daltons)	No. of Cistrons	Total no. of rare alleles	μ per cistron (x10⁶)	
								Kimura & Ohta's method	Nei's method
Non-polymorphic	<0.02	Multimers	0.0003	2372	37,000	11	5	7.28	3.22
		Monomers	-	2322	34,000	3	-	-	-
		All	0.0002	2361	36,357	14	5	5.72	2.53
Polymorphic	0.02-0.10	Multimers	0.0380	1290	24,000	1	-	-	-
		Monomers	0.0370	3322	37,167	6	8	21.37	5.43
		All	0.0371	3032	35,286	7	8	18.32	4.75
	0.10-0.30	Multimers	0.1903	2327	43,333	3	1	5.35	2.37
		Monomers	0.2097	3919	51,000	1	2	33.75	12.52
		All	0.1951	2725	45,250	4	3	12.03	5.13
	All	Multimers	0.1522	2068	38,500	4	1	4.01	1.84
		Monomers	0.0616	3408	39,143	7	10	22.90	6.47
		All	0.0946	2920	38,909	11	11	16.04	4.89

TABLE 4.5 SUBUNIT SIZE AND ELECTROMORPH MUTATION RATES

Range of subunit size	Type of enzyme	No. of cistrons	No. of rare alleles	Mean subunit size (in daltons)	Mean sample size	μ per cistron ($x10^6$)	
						Kimura & Ohta's method	Nei's method
<25,000 daltons	Multimers	4	-	17,750	2117	-	-
	Monomers	2	1	18,500	4016	8.02	3.12
	All	6	1	18,000	3776	2.72	1.06
25,000-50,000	Multimers	8	3	43,375	2410	6.08	2.65
	Monomers	5	3	34,400	2474	9.63	4.20
	All	13	6	39,923	2434	7.38	3.25
>50,000 daltons	Multimers	3	3	55,667	2206	16.04	7.21
	Monomers	3	6	55,667	3633	32.08	6.37
	All	6	9	55,667	2919	24.06	6.72

multimers and monomers. The pattern in the three categories is of systematic increase with larger subunit sizes. Multimers have consistently lower mutation rates as compared with monomers although the mean value of subunit sizes and sample sizes are similar.

4.2.3. Discussion.

From the observations outlined, it is obvious that the structural constraints and cistron sizes of enzymes, besides the role of sample size, determine to a large extent the relative magnitudes of electromorph mutation rates. Any comprehensive estimate of mutation rates for a population will thus have to be weighted for sample size and subunit size. In the present data, weighting by these factors leads to a general reduction in the average mutation rates because of the higher invariant nature of loci with low sample sizes and subunit sizes (table 4.6). Adjustment for amino-acid residues involved in surface interactions in multimers reduces further the average mutation rates. This gives new estimates for μ per cistron per generation in Australian Aborigines as 6.19×10^{-6} and 2.35×10^{-6} by the methods of Kimura and Ohta (1969) and Nei (1977) respectively. The differences between monomers and multimers are increased substantially after these modifications.

TABLE 4.6 ELECTROMORPH MUTATION RATES ($x10^6$) IN AUSTRALIAN ABORIGINES WEIGHTED FOR SAMPLE SIZE, SUBUNIT SIZE AND PROPORTION OF CISTRON INVOLVED IN SURFACE INTERACTIONS

Weighted by	Kimura and Ohta's Method			Nei's Method		
	Multimerics	Monomerics	Total	Multimerics	Monomerics	Total
Unweighted	6.42	16.03	10.26	3.33	4.64	3.58
Sample size	3.36	19.27	8.67	1.73	3.97	3.28
Sample size + subunit size	1.68	22.31	6.67	0.86	4.61	2.51
Sample zize + subunit size + molecular surface interactions	1.39	22.31	6.19	0.71	4.61	2.35

In principle the interlocus variability in the mutation rates arising from the various cistron sizes should be minimized if we calculate the mutation rates per base pair per generation rather than per cistron per generation. The estimates of μ per base pair per generation are given in table 4.7.

Despite the incorporation of modifications necessitated by the physico-chemical constraints of the molecules and sample sizes, differences among mutation rates still exist. For example, the relationship between the subunit size and mutation rates does not resolve into a simple linear function. Similarly, the differences between the multimeric and monomeric enzymes are increased when adjustments are made for the variation arising from the sample size and subunit size, yet the distinction between polymorphic, monomeric, large-subunit-enzymes and monomorphic, multimeric, small-subunit-enzymes is clear cut. This indicates that while making comparisons for electromorph mutation rates among various human populations, the number and type of loci included in the estimations should be taken into account.

TABLE 4.7 UNADJUSTED ELECTROMOPH MUTATION RATES PER BASE PAIR IN AUSTRALIAN ABORIGINES

Type of Enzyme	μ per base pair ($x10^{8}$)	
	Kimura and Ohta's method	Nei's method
Multimers	1.71	0.87
Monomers	8.25	1.70
Total	4.11	1.56

4.3. Hypergeometric Sampling and Estimation of Mutation Rate.

Recently Rothman and Adams (1978) have presented a statistical approach for estimating the average number of variants per locus in a population of 2N genes (K) from the number per locus in a sample (k) using the binomial approximation to the hypergeometric probability distribution. The estimated value of K is then used to estimate indirectly the mutation rate from the number of singletons extant in the population K_s or Kg(1) where g(1) is the relative frequency of singletons in the population.

Using a binomial sampling equation, Chakraborty (1981) has given the conditional expectation of the number of rare alleles in a sample (k_r) of 2n genes, where the total population size N is known, under the infinite alleles model of Kimura and Crow (1964). Using the method of moments estimation approach he has suggested the estimation of mutation rate indirectly from the observed number of rare alleles in the sample ($\hat{k}_r$).

Neel and Rothman (1978) and Chakraborty (1981) have used the algorithms of Rothman and Adams (1978) and Chakraborty (1981) respectively to generate the estimates of mutation rate in 12 Amerindian tribes. A comparison of the two sets of estimates reveals some interesting similarities as well as differences.

1. The estimates of μ by Rothman and Adams' method (μ_{RA}) are on an average 59% higher than those obtained by Chakraborty's method (μ_{CHAK}). These range from 0.71 to 13.83 fold .
2. The coefficient of correlation between the two sets of estimates is highly significant ($r=0.844$; $P<10^{-3}$).
3. While the estimates by Chakraborty's method show a significant negative correlation with population size ($r=0.583$; $P=0.023$), Rothman and Adams' method yields nonsignificant values. However, no correlation is observed with sample size for both the sets of estimates.
5. The correlation between the sampling fraction, $f=n/N$, and the ratio of mutation rates, $\hat{\mu}_{CHAK}/\hat{\mu}_{RA}$ is highly significant ($r=0.743$; $P<10^{-3}$). This observation leads to interesting intrapolation. For sampling fraction (f) of less than 5%, this ratio is 0.113; for the value of f between 0.30 and 0.40 the ratio becomes 0.74.

From the above observations it is obvious that the estimation procedures of Rothman and Adams (1978) and Chakraborty (1981) lead to significantly different results, the differences among which are masked by averaging over a set of data. Three major points of difference between the two procedures can be recognized:

1. The binomial approximations to the hypergeometric probability distributions are made differently,
2. the two methods use different aspects of the data; while the Rothman and Adams' method uses the singletons in the population after extrapolation, the method of Chakraborty utilizes only those rare alleles which are recovered in the sample, and
3. the two methods use different models. While Rothman and Adams use the equilibrium equation between the number of singletons gained and lost from the population for their procedure, Chakraborty uses the diffusion approximation of the infinite alleles model.

It may thus be important to find out the relative contribution of these three factors in producing the observed differences in the estimates of mutation rate. Using the data of Neel and Rothman (1978) and Neel (1978b) on 12 Amerindian tribes I have tried to illustrate the role of one or more of the above points in determining the differences between the estimates.

4.3.1. Formulations.

The extrapolation of the sample number of different alleles recovered in a sample (k) to its population value (K) was first suggested by Ewens (1972) using the Wright-Fisher infinite

alleles model of Kimura and Crow (1964) as

$$K = k + \int_{2N^{-1}}^{2n^{-1}} \Phi(x)dx \qquad (4.1)$$

where $\Phi(x)dx$ defines the frequency spectrum having the property that $\Phi(x)dx$ is the mean number of alleles in the population with frequency in $(x,\Delta x+x)$. The expression $\Phi(x)$ is given as

$$\Phi(x)dx = \theta x^{-1}(1-x)^{\theta-1}dx$$

The equation (4.1) solves approximately, as

$$K-k = \theta\log(N/n)-\theta(\theta-1)[2n^{-1}-2N^{-1}] \qquad (4.2)$$

where N is the effective population size (taken generally as the number of individuals in the reproductive age group, 15-44 yrs), n is the number of individuals sampled (or 2n genes) and $\theta=4N\mu$ is the scaled mutation rate. The equation (4.2) is rewritten in quadratic, as

$$K-k \simeq \theta A' - \theta(\theta-1)B'$$

$$= \theta(A'+B') - \theta^2 B' \qquad (4.3)$$

where $A' = \log(N/n)$ and $B' = [2n^{-1}-2N^{-1}]$

The method of moment yields an estimator of θ, $\hat{\theta}_{K-k}$, given by solving the root of the quadratic in (4.3) as

$$\theta_{K\text{-}k} = \frac{(A'+B') - \sqrt{(A'+B')^2 - 4B\hat{K}\text{-}k}}{2B'}, \qquad (4.4)$$

the other root being inadmissible.

The solution (4.4) is, however, unattractive in that there are two, rather than one, unknowns in θ and K. An alternative approach was taken by Nei (1977) who equated the number of alleles segregating in the population within the frequency range $(2n^{-1}, q)$, where the upper limit of q is taken arbitrarily as 0.01 or 0.05, for the number of rare alleles (q) recovered in the sample (k_r). Accordingly, for the infinite alleles model, Nei's approach yields

$$\begin{aligned} k_r &= \int_{2n^{-1}}^{q} \Phi(x)dx \\ &\simeq \theta \log(2nq) - \theta(\theta-1)(q-2n^{-1}) \\ &= \theta\{\log[2nq]+q-2n^{-1}\} - \theta^2(q-2n^{-1}) \end{aligned} \qquad (4.5)$$

which is quite close to his sampling equation derived by using the infinite sites model as

$$k_r = \theta \log (2nq)$$

It is also obvious, however, that for $2n>1/q$ the equation (4.4) *de facto* utilizes only rare alleles. Besides, equation (4.4) utilizes that part of the information on the rare alleles in the population which are not detected in the sample. It is

expected, therefore, that equations (4.4) and (4.5) should yield almost similar results for $2n>1/q$.

To estimate K from the number of alleles in the sample, the extrapolation can be obtained by using the binomial approximation to the hypergeometric probability (Rothman and Adams, 1978). Following Lieberman and Owen (1961) the approximation has a simple solution, i.e.

$$K = k\,(1-Z)^{-1} \qquad (4.6)$$

where

$$Z = \Sigma g(j)\left(1- \frac{2n}{2N- \frac{j-1}{2}}\right)^{j}$$

given that g(j) is the relative proportion of alleles represented by j copies in the population. Substituting in (4.3), we get

$$\hat{k}\, Z(1-Z)^{-1} = K-k$$

$$= \theta(A'+B') - \theta^2 B' \qquad (4.7)$$

A different approach is taken by Chakraborty (1981) to incorporate the sampling effects in the infinite alleles model. His binomial sampling equation is given as:

$$k_r = \sum_{j=1}^{[2nq]} \binom{2n}{j} \theta \int_0^1 x^{j-1}\,(1-x)^{\theta+2n-j-1} dx$$

$$\simeq \theta A - \theta^2 B \qquad (4.8)$$

where $A = \sum_{j=1}^{[2nq]} j^{-1}$ and $B = \sum_{j=1}^{[2nq]} (2n-j)^{-1}$

I shall, however, use a slightly different form of (4.8), the justification for which is given in appendix A. However, this modified approach will still be referred to as Chakraborty's method in the text following. Accordingly

$$\hat{k}_r \simeq \theta(A+D) - \theta^2 D \qquad (4.9)$$

where A is as defined in equation (4.8) and D is given as

$$D \simeq \frac{2nq}{2n-1} + \frac{2nq(2nq+1)}{4(2n-1)^2}$$

Using Chakraborty's (1981) sampling approach an estimate for the number of rare alleles in the population, not recovered in the sample, is given as

$$K_r - k_r \simeq \theta \log(N/n) \qquad (4.10)$$

which is quite similar to (4.3).

For single copy alleles in the sample (k_s) the equation (4.8) has an exact solution, i.e.

$$\hat{k}_s = 2n\theta/(2n+\theta-1) \qquad (4.11)$$

which yields an estimator of θ as

$$\hat{\theta}_{k_s} = (2n-1)\hat{k}_s/(2n-\hat{k}_s) \qquad (4.12)$$

For singletons in the population K_s, Rothman and Adams (1978) have given an estimator of mutation rate, $\hat{\mu}_{RA}$ as

$$\hat{\mu}_{RA} = \frac{K}{2N} [\tilde{g}(1) - \sum_{j=1}^{2N} g(j)P_{j1}] \qquad (4.13)$$

where $\tilde{g}(j)$ is the estimated relative proportion of alleles in the jth allelic state given that j is the number of copies by which an allele is represented in the population and P_{j1} is the probability that an allele represented by j copies in the previous generation is represented now by singletons only. The estimator for k is described already by equation (4.6).

4.3.2. Results and discussion.

To illustrate the differences in the estimates of mutation rate generated by using various estimation procedures or different aspects of allelic data as input, the data on 12 Amerindian tribes given by Neel and Rothman (1978) and Neel (1978b) have been utilized. The summary of various statistics used is shown in table 4.8. Since the number of variants detected in a sample decreases with more conservative upper frequency limits, there are comparatively more observations available by the extrapolation approach of Rothman and Adams, than by the rare allele approach of Chakraborty (1981) or single copy approach indicated in this study.

TABLE 4.8 SUMMARY OF THE NUMBER OF VARIANTS DETECTED PER POLYPEPTIDE IN 12 AMERINDIAN TRIBES (BASED ON DATA IN NEEL AND ROTHMAN, 1978 AND NEEL, 1978b)

Tribe	Effective population size (x2)	Average No. of genes Sampled	Alleles sampled			
	$2N_e$	2n	$\hat{k}_s$	$\hat{k}_r$	$\hat{k}_t$	$K-\hat{k}$**
Ayoreo	1,440	194	-	-	-	-
Baniwa	1,440	362	0.0370	0.0370	0.0741	0.0674
Cayapo	1,440	524	-	0.0357	0.0714	0.0412
Guayumi	28,800	466	0.0370	0.0370	0.0741	0.0734
Kraho	576	184	-	0.0357	0.0357	0.0244
Macushi	3,840	480	0.0370	0.0741	0.1111	0.2119
Makiritare	1,440	496	-	0.0741	0.0741	0.0459
Panoa	17,280	320	-	-	0.0370	0.4704
Piaroa	2,880	140	-	0.0417	0.0417	0.2031
Wapishana	1,920	590	0.0714	0.1071	0.1786	0.1280
Xavanate	1,632	312	0.0769	0.1538	0.1538	0.1892
Yanomama	14,400	1,920	-	-	0.0357	0.0638

* k_s, singletons; k_r, rare alleles; k_t, all private alleles

** Number of alleles in the population NOT included in the sample

These differences become more significant when the data are on small sized populations with comparatively more private polymorphisms.

Four different estimates of mutation rate for each of the 12 Amerindian tribes are shown in table 4.9. The estimation procedures used are those of Rothman and Adams ($\hat{\mu}_{RA}$), Chakraborty ($\hat{\mu}_{CHAK}$), singleton approach (μ_{k_s}) and the Ewens' approximation equation ($\hat{\mu}_{K-k}$). In addition, estimates utilizing Chakraborty's method is extended to all the alleles detected in the sample ($\hat{\mu}_{k_t}$).

With singletons as input the estimates of μ_{k_s} and μ_{RA} exhibit more than three-fold difference in the average estimates (0.50×10^{-5} per locus/generation against 1.71×10^{-5} per locus/generation). These differences are due partly to the lack of recovery of singletons in certain tribes. For five tribes with non-zero values of k_s, the average estimates become respectively 1.20×10^{-5} per locus/generation and 2.59×10^{-5} per locus/generation. The average estimates do not show much variation especially when we realize that the extrapolation approach of Rothman and Adams leads to more assured recovery of $\hat{K}_s$ values. Nine of the twelve tribes yield estimates of $\hat{\mu}_{CHAK}$ with an average over all the tribes of 0.78×10^{-5} per locus/generation. In comparison the estimates for the $\hat{\mu}_{K-k}$ exhibit an average value of 1.36×10^{-5} per locus/

TABLE 4.9 ESTIMATES OF MUTATION RATE, μ ($\times 10^5$) FOR TWELVE AMERINDIAN TRIBES OBTAINED BY USING DIFFERENT ESTIMATION PROCEDURES

	Tribe	Chakraborty's binomial approximation			Rothman and Adams binomial approximation	
		$\hat{\mu}_{k_s}$	$\hat{\mu}_{k_r}$	$\hat{\mu}_{k_t}$	$\hat{\mu}_{R-A}$	$\hat{\mu}_{K-k}$
1.	Ayoreo	-	-	-	-	-
2.	Baniwa	1.28	0.70	0.44	2.06	1.69
3.	Cayapo	-	0.55	0.41	1.64	1.41
4.	Guayumi	0.06	0.03	0.02	0.83	0.45
5.	Kraho	-	2.06	0.60	2.19	1.85
6.	Macushi	0.48	0.46	0.24	1.76	1.33
7.	Makiritare	-	1.13	0.42	1.75	1.49
8.	Panoa	-	-	0.02	0.61	0.34
9.	Piaroa	-	0.72	0.15	1.78	1.16
10.	Wapishana	1.85	1.13	0.75	3.35	2.82
11.	Xavanate	2.35	2.57	0.84	4.40	3.50
12.	Yanomama	0.50	0.78	0.33	1.71	1.36
	Average	0.50	0.78	0.33	1.71	1.36

generation. The respective averages for tribes with non-zero values of $\hat{k}_r$ are 1.04×10^{-5} per locus/generation and 1.76×10^{-5} per locus/generation.

Two types of methodological differences are discerned between the estimators $\hat{\mu}_{R-A}$ and $\hat{\mu}_{K-k}$. These are the use of different aspects of allelic data and the choice of different mutation models. Since Neel and Rothman (1978) used the same value (0.2096) of $g(1) - \Sigma g(j)P_{j1}$ for all the tribes in their calculations, it is only to be expected that the proportion of $\hat{\mu}_{K-k}/\hat{\mu}_{R-A}$ should be more or less similar over all the tribes. We observe, however, this proportion to vary from 0.54 - 0.86 with an average value of 0.70. While the deviation from unity indicates that the choice of singletons yields higher values of $\hat{\mu}_{R-A}$, the variation in the ratio $\hat{\mu}_{K-k}/\hat{\mu}_{R-A}$ is unexplainable. It may, however, be noted that this mutation rate ratio is positively correlated with the sampling fraction ($r=0.895$; $P<0.001$).

A comparison of the role of singletons vs. rare alleles in raising the estimates of μ is also made for the binomial approximations given by Chakraborty (1981). For five tribes with non-zero values of k_s, the mean mutation rates for $\hat{\mu}_{k_s}$ and $\hat{\mu}_{CHAK}$ are 1.20×10^{-5} per locus/generation and 0.98×10^{-5} per locus/generation, respectively which indicates that the choice of singletons does yield higher

estimates in comparison with use of rare alleles. Similarly, for the nine tribes with positive k_r, the value of μ_{k_r} (0.98×10^{-5} per locus/generation) is higher than that obtained by using all alleles, μ_{k_t} (0.43×10^{-5} per locus/generation).

The ratio $\hat{\mu}_{CHAK}/\hat{\mu}_{RA}$, however, shows much wider range than the ratio $\hat{\mu}_{K\text{-}k}/\hat{\mu}_{RA}$. The respective ranges with the average values are 0.04-0.94 (0.47) and 0.54-0.84 (0.69). While some of the differences between $\hat{\mu}_{CHAK}$ and $\hat{\mu}_{RA}$ can be atrributed to the initial advantages in using the extrapolation approach of Rothman and Adams (1978), there still exist significant differences between the two estimates for individual populations.

The differences produced by the choice of the branching process model by Rothman and Adams and the diffusion approximation approach by Chakraborty can be explained, as follows:

If we approximate Chakraborty's equation,

$$\sum_{j=1}^{[2nq]} k_j \simeq \theta \sum_{j=1}^{[2nq]} j^{-1} - \theta^2 \sum_{j=1}^{[2nq]} (2n-j)^{-1}$$

for small θ, to

$$\sum_{j=1}^{[2nq]} k_j \simeq \theta \sum_{j=1}^{[2nq]} j^{-1} \qquad (4.13)$$

Then θ can be replaced by k_1 as

$$\Sigma k_j = k_1 \Sigma j^{-1} \qquad (4.14)$$

By scaling $k_1 = 1$ and assuming that the relative frequencies of various allelic states are the same in both the sample and the population, we expect

$$\Sigma \hat{k}_j = \Sigma j^{-1}$$

where $\hat{k}_j$ is the observed scaled value of k_j. For Yanomama tribes, by using the branching process argument, Rothman and Adams (1978) have provided the values of k_j. Summing over the first ten terms, the two values are

$$\sum_{j=1}^{10} \hat{k}_j = 2.086$$

$$\sum_{j=1}^{10} j^{-1} = 2.929$$

which indicates that the estimates of μ by Rothman and Adams' method will be on an average 40% higher than that of Chakraborty. These differences are, however, systematic and do not indicate the cause of the presence of outliers nor the significant correlations of μ_{CHAK}/μ_{RA} with the sampling fractions.

One of the main reasons for the presence of outliers could be the choice of different approach-

es to the approximations of the hypergeometric probability. It is known that the binomial approximation to the hypergeometric sampling becomes poorer with rise in the sampling fraction. By virtue of the symmetry of the hypergeometric probability, this applies to the allelic frequencies too. Besides, to obtain best approximation for a particular value of the sampling fraction and the allelic frequency one should use the smallest value of the cumulative binomial probability distribution, denoted by E(a,b,p). This smallest value is realized when a is the smallest (Lieberman and Owen, 1961).

When the approximation is made over the whole range of allelic frequency, b (0,1), for a fixed value of the sampling fraction, f, there are two stages in approximations, i.e. when $f<p$ and when $f>p$. By using the same approximation over the whole range of p, one obtains poor approximation to the hypergeometric; the extent of this overestimation of E (a, b, p) being related to the value of f.

Both Rothman and Adams (1978) and Chakraborty (1981) have made these approximations in their sampling algorithms over the full range of p. Their approaches are, however, directly opposed to each other. While Rothman and Adams use E (p,b,f) in their formulation, Chakraborty (1978) utilizes

E(f,b,p). For small f, Chakraborty's approach admits less error than that of Rothman and Adams. For large F, it is *vice versa*. These poorer approximations in turn produce outliers for small values of f. Since in the data of Neel and Rothman (1978) the highest value of f is 0.364, there are few outliers observed for higher values of f. These differences are seen in more than an order of magnitude difference in μ for f about 2% and a spurious significant correlation between μ_{CHAK}/μ_{R-A} and f produced by the lack of values for higher ranges of f. However, it is not clear if these differences are accentuated by the smaller values of b, which are 2n and 2nq respectively by the two methods.

These observations lead to the question as to which method is more appropriate and whether the estimates generated by one method are over-estimates or under-estimates. The results indicate the desirability of choosing a method which reduces the number of outliers in the estimates. The use of equation (4.4) and (4.7) may provide such estimates.

Chapter 5

RELATIVE ELECTROMORPH MUTATION RATES

5.1. Introduction.

In the previous chapter, attention was drawn to the positive correlation between electromorph mutation rate and subunit molecular weight using data for Aboriginal populations in Australia. The same data was used to show a negative correlation with the number of subunits in the functional enzyme and also to illustrate the effect of sample size on the ability to detect electromorphs in the population. The analysis of electromorph mutation rates has now been extended to include data available from intensive surveys carried out by several different investigators for a number of major human populations: a total of more than 800,000 single locus tests has been analysed.

Two different stragegies have been employed in examining the factors influencing electromorph mutation rates. In the first, the relationships of sample heterozygosities, or mean single locus heterzygosities over a set of related populations, are analyzed, using both parametric and non-parametric correlation methods. In the second, the analysis is restricted simply to the relationship between the number of different electrophoretic alleles and the size and structure of protein molecules.

Using heterozygosity as a measure of genetic

variability the dependence of neutral mutation rates on subunit molecular weight was demonstrated by Brown and Langley (1979), Turner *et al.* (1979), Ward (1978) and Koehn and Eanes (1977, 1978) for various vertebrate and invertebrate populations. This class of relationship, however, was not demonstrated in single species tests of *Colias* (Johnson, 1979), *Drosphila* (Johnson, 1979; Voelker *et al.*, 1980b) and man (Harris *et al.*, 1977; Nei *et al.*, 1978; Bhatia, 1980). However, in single species tests, Harris *et al.*, (1977), Ward (1977) and Bhatia (1980) have shown heterozygosity to be negatively correltated with subunit numbers.

Using the second strategy, a relationship between the average number of different alleles per locus and subunit size was demonstrated by Eanes and Koehn (1978) and Bhatia (1980) in pooled data on human populations and Australian Aborigines respectively. This class of relationship is, however, difficult to evaluate, as the non-parametric estimates of the number of electrophoretic alleles depend critically upon sample size (Nei, 1977; Eanes and Koehn, 1978; Rothman and Adams, 1978; Bhatia, 1980) and upon the experimental techniques employed to discriminate allelic variants (Johnson, 1977).

Variability in the estimates of mutation rate from protein data, corresponding to the variation in subunit size, has been shown to follow the gamma

distribution (Nei, *et al.*, 1976b; Fuerst *et al.*, 1977; Zouros, 1979). Zouros (1979) has used these relationships to generate estimates of relative electromorph mutation rates (REMR) in various natural populations. Using total heterozygosity as input, he found the REMRs to vary more than 500 times over a set of protein loci.

Because of the lack of correlation between heterozygosity and subunit size, extension of Zouros' approach to human data will have only a limited value. Instead the data on rare allele variability may be used to generate relative estimates of mutation rate because of its known dependence on subunit molecular weights. In the present chapter, therefore, rare allele variability, expressed both as rare allele heterozygosity as well as the number of rare alleles, is utilized to estimate the REMRs.

5.2. The Laboratory Data

In this analysis data on population surveys for electrophoretic variants in 10 major ethnic groups have been included. The surveys on Australian Aborigines, Melanesians, Iranians and South Asian tribal populations are from published and unpublished sources of data in this laboratory. The surveys adopted from other sources are: Amerindians (Neel, 1978b), Japanese (Neel *et al.*, 1978); GPT data from Ishimoto and Kuwata, 1974);

English (as compiled by Neel *et al*, 1978; Welch *et al*, 1975 for GPT data); Aymara Indians (Schull *et al*, 1978); South African Khoisan and Negroid populations (based on work by Professor T. Jenkins and his collaborators and compiled by Bhatia *et al.* in preparation).

The data on Melanesians is subdivided into two linguistic groups, namely Austronesians and non-Austronesians, because of their different origins (Wurm, 1975a). Rare alleles, assigned on the basis of higher frequency to one language group, have been excluded from the other.

The data have been compiled for 27 protein loci (17 multimers and 10 monomers) and are listed in table 5.1. The multimeric loci are all dimers except the two LDH loci which are tetramers. Subunit molecular weights are taken from the tabulations of Darnall and Klotz (1975) and Hopkinson, *et al.*, (1977).

A rare allele has been designated here as one with less than 20 copies in 1,000 determinations. For each population a separate list of rare alleles was prepared. Rare allele heterozygosity (H_r) is defined here as the number of copies contributed by rare alleles/1,000 determinations. The second parameter, the number of rare alleles (K_r) is simply a count of different rare alleles recovered at each locus. For some purposes K_r is specified per 1,000 determinations.

TABLE 5.1 INTERLOCUS VARIABILITY IN THE FREQUENCY OF RARE ALLELES AND ESTIMATES OF RELATIVE ELECTROMORPH MUTATION RATES (REMR)

LOCUS	No. of determinations	Rare alleles		Rare allele heterozygosity (H_r)	No. of different rare alleles (K_r)	Relative electromorph mutation rates (REMR)	
		Number	Copies			REMR(1)	REMR(2)
	A	B	C	$D=\frac{C}{A} \times 1000$	$E=\frac{B}{A} \times 1000$	$F=\frac{D_i}{\Sigma D_i}$	$G=\frac{E_i}{\Sigma E_i}$
				MULTIMERS			
Hb-α	49,191	11	170	3.46	0.224	0.0477	0.0256
Hb-β	49,191	11	39	0.79	0.224	0.0109	0.0256
SOD_A	30,327	2	11	0.36	0.066	0.0050	0.0076
GLO	5,658	0	0	0.00	0.000	0.0000	0.0000
EsD	18,993	3	4	0.21	0.158	0.0029	0.0181
MDH	33,186	8	153	4.61	0.241	0.0635	0.0277
LDH_A	34,886	13	47	1.35	0.373	0.0186	0.0427
LDH_B	34,886	8	71	2.04	0.229	0.0281	0.0262
Hp	38,563	8	9	0.23	0.207	0.0032	0.0237
GOT	8,352	4	4	0.48	0.479	0.0066	0.0548
Pep A	32,853	15	59	1.81	0.457	0.0250	0.0523
ICD_s	21,994	9	14	0.64	0.409	0.0088	0.0468
Pep D	7,669	4	49	6.39	0.522	0.0880	0.0597
GPT	14,769	7	23	1.56	0.474	0.0215	0.0542
6PGD	46,884	18	188	4.01	0.384	0.0552	0.0439
Cp	23,244	14	115	4.95	0.602	0.0682	0.0689
PHI	28,060	25	103	3.67	0.891	0.0505	0.1020
				MONOMERS			
ACP_1	46,855	7	45	0.96	0.149	0.0133	0.0171
AK_1	38,385	2	11	0.29	0.052	0.0040	0.0060
CA_1	26,889	5	15	0.56	0.186	0.0077	0.0213
CA_2	17,502	2	40	2.29	0.114	0.0316	0.0130
PGK	17,700	2	112	6.33	0.113	0.0872	0.0129
PGM_1	49,605	27	96	1.94	0.544	0.0267	0.0623
Pep B	33,310	15	118	3.54	0.450	0.0488	0.0515
PGM_2	48,550	14	304	6.26	0.288	0.0862	0.0330
Alb	30,264	9	276	9.12	0.297	0.1256	0.0340
Tf	38,091	23	192	5.04	0.604	0.0694	0.0691

The relationships between rare allele variability and molecular structure were tested using linear regression methods. Whenever necessary, the variables were log transformed to equalize and normalize the variances. Both Pearson's product moment and Spearman's rank order correlations were computed to test the correspondence between different variables.

5.3. Results.

Table 5.1 shows the distribution of the total number of rare alleles (B) and total number of copies (C) and sample sizes (n), for the 27 protein loci. Columns D and E of the table show the observed estimates of rare allele heterozygosity (H_r) and number of rare alleles (K_r) per 1000 determinations respectively. The weighted mean subunit sizes, sample sizes and rare allele heterozygosities for various classes of subunit size are shown in table 5.2.

5.3.1. Interlocus variability.

5.3.1.1. Number of rare alleles (K_r).

A total of 266 different rare alleles, with an average recovery of 1 rare allele for every 3,016 determinations, was detected. The range is from none in 5,658 determinations for glyoxalase (GLO) to one in 1,122 determinations for phosphohexose isomerase (PHI). Despite a singificant

TABLE 5.2 SUBUNIT SIZE, QUATERNARY STRUCTURE AND RARE ALLELE VARIATION IN 12 HUMAN POPULATIONS

Range of subunit size	Type of protein	No. of cistrons	Rare alleles		Mean subunit size	Mean sample size	Rare allele heterozygosity (per 1000 determinations	No. of rare alleles (per locus)	No. of rare alleles (per 1000 determinations
			No.	copies					
<25,000 daltons	Multimers	4	24	220	17,750	33,592	1.637	6.00	0.179
	Monomers	2	9	56	18,500	42,620	0.657	4.50	0.106
	Total	6	33	276	18,000	36,601	1.256	5.50	0.150
25,000 -50,000 daltons	Multimers	10	79	433	41,300	24,609	1.760	7.90	0.321
	Monomers	3	9	167	36,000	20,697	2.690	3.00	0.145
	Total	13	88	600	40,076	23,706	1.946	6.77	0.286
>50,000 daltons	Multimers	3	57	406	55,667	32,279	4.135	19.00	0.580
	Monomers	5	88	986	65,200	39,964	4.934	17.60	0.440
	Total	8	145	1,312	61,625	37,251	4.671	18.12	0.487
Total	Multimers	17	160	1,059	38,294	28,156	2.212	9.41	0.334
	Monomers	10	106	1,209	47,100	34,715	3.482	10.60	0.305
	Total	27	266	2,268	41,555	30,585	2.746	9.85	0.322

correlation between the recovery of rare alleles and sample size ($r=0.544$; $P<0.002$), sampling error is unlikely to explain the failure to recover variants for glyoxalase (GLO). The possibility of testing 5,658 individuals without detecting a variant is very low ($P<0.001$).

The mean unweighted number of rare alleles ($\bar{K}_r$) per 1,000 determinations in monomers and multimers are 0.279 ± 0.088 and 0.349 ± 0.053 respectively. The difference is statistically insignificant, thereby discounting the role of quaternary structure in introducing new alleles.

There is a significant positive correlation between the number of different rare alleles (K_r) and subunit size (m). The values of r_{Km} for total, multimeric and monomeric loci are shown in table 5.3. Only 34% of the variability in the number of different rare alleles is explained by variability in subunit size. This proportion rises to 55% when the partial correlations are computed, after controlling for sample size. Considered separately, both multimers and monomers show better correspondence with their respective molecular weights (see table 5.3). The value of r^2 for multimers and monomers is increased to 74% and 57% respectively, when adjustments are made for sample size as control variable. The estimated parameters for the regression line $y = a + bX$ are: $\hat{a} = 0.02402$ and

TABLE 5.3 CORRELATION COEFFICIENTS (r) BETWEEN MOLECULAR WEIGHT, SAMPLE SIZE AND PARAMETERS OF RARE ALLELES AND THE PROPORTIONS OF VARIANCE EXPLAINED BY MOLECULAR WEIGHT VARIATION (r^2).

Parameter	Type of Protein	Sample size		Subunit size	
		r	r^2	r	r^2
Number of rare alleles (K_r)	Multimers	0.5149*	0.2651	0.5118*†	0.2619
	Monomers	0.6269*	0.3930	0.6402*†	0.4099
	Total	0.5441**	0.2960	0.5834***†	0.3403
Rare allele heterozygosity (H_r)	Multimers	0.0678	0.0046	0.4314*	0.1861
	Monomers	(-)0.1652	0.0273	0.7511**	0.5641
	Total	0.0462	0.0021	0.6411**	0.4110

* $0.01<P<0.05$

** $0.001<P<0.01$

*** $P<0.001$

† Partial correlations after controlling for sample size are 0.8450^{***} 0.7590^{***} and 0.7434^{***} for multimers, monomers and total proteins respectively.

$\hat{b}$ = 0.00023. The small value of $\hat{b}$ is due to the units used for expressing molecular weights. The scattergram for the values at each locus is shown in Fig. 5.1.

5.3.1.2. Rare allele heterozygosity (H_r).

The estimates of rare allele heterozygosity (H_r) are not related to fluctuations in sample size ($r=0.046$; $P>0.410$) or to the number of different rare alleles ($r=0.272$; $P>0.085$) (the rank order correlations for the latter are, however, significant). However, a significant positive correlation does exist with subunit size ($r=0.641$; $P>0.001$) with r^2 explaining more than 41% variability contributed by molecular weight. These results are specially significant in view of the lack of correlation between total heterozygosity and molecular weight in human populations. Both multimers and monomers similarly exhibit significant correlations, although the value of r^2 in multimers is only 18% against 56% for monomers (see table 5.3).

The unweighted mean values of rare allele heterozygosity ($\bar{H}_r$) 2.70 ± 0.47, 2.15 ± 0.48 and 3.63 ± 0.93 for total, multimeric and monomeric loci respectively. In contrast with the results derived from the number of rare alleles (K_r) per 1,000 determinations, the results for rare allele heterozygosity exhibit significant differences

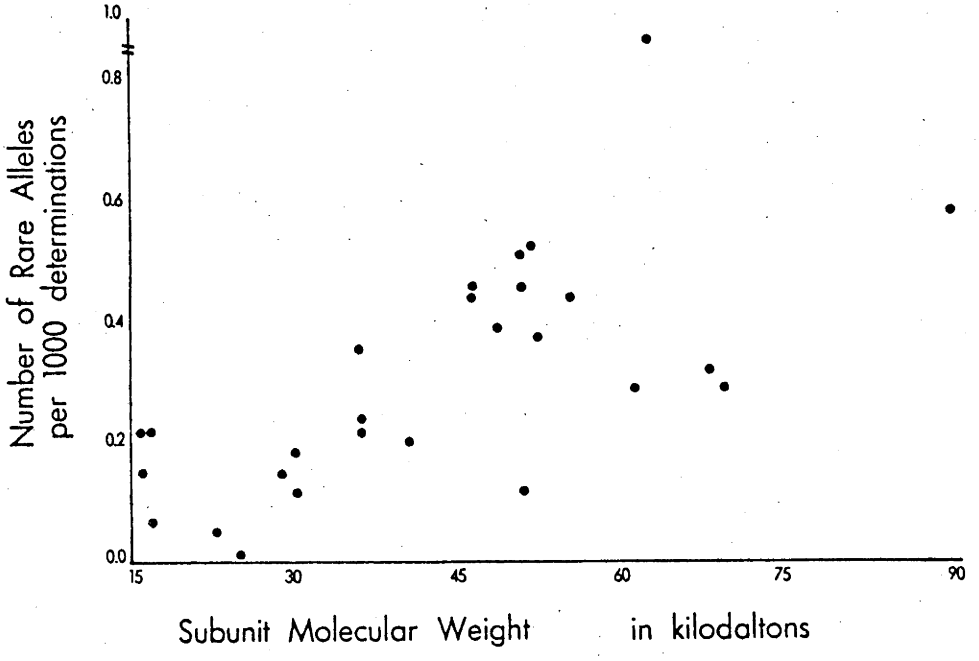

Fig. 5.1 Relationship between the number of different rare alleles per 1,000 determinations at a locus and the respective subunit molecular weights in human populations.

between monomers and multimers. The role of molecular constraints present in multimers in reducing genetic variability is discernible in this parameter.

The scattergram for the values of rare allele heterozygosity (H_r) and subunit molecular weight at each locus is shown in Fig. 5.2. The linear regression is

$$\hat{y} = -0.84998 + 0.00009X$$

5.3.1.3. Relative electromorph mutations rates (REMR).

Two different estimates of relative electromorph mutation rates (REMR) were obtained.

REMR(1) represents the scaled value of rare allele heterozygosity, so that for any locus

$$REMR(1) = \frac{H_r}{\Sigma H_r}$$

and REMR(2) represents the scaled value for the number of different rare alleles, so that

$$REMR(2) = \frac{K_r}{\Sigma K_r}$$

The values of REMR(1) and REMR(2) are given in the last two columns of table 5.1. Although both the estimates of REMR show positive correlations with subunit molecular weight and have similar rankings, the variability exhibited by the two methods differs widely. After excluding the invariant locus (GLO), the ratio between the lower and upper values for REMR(1) and REMR(2) is

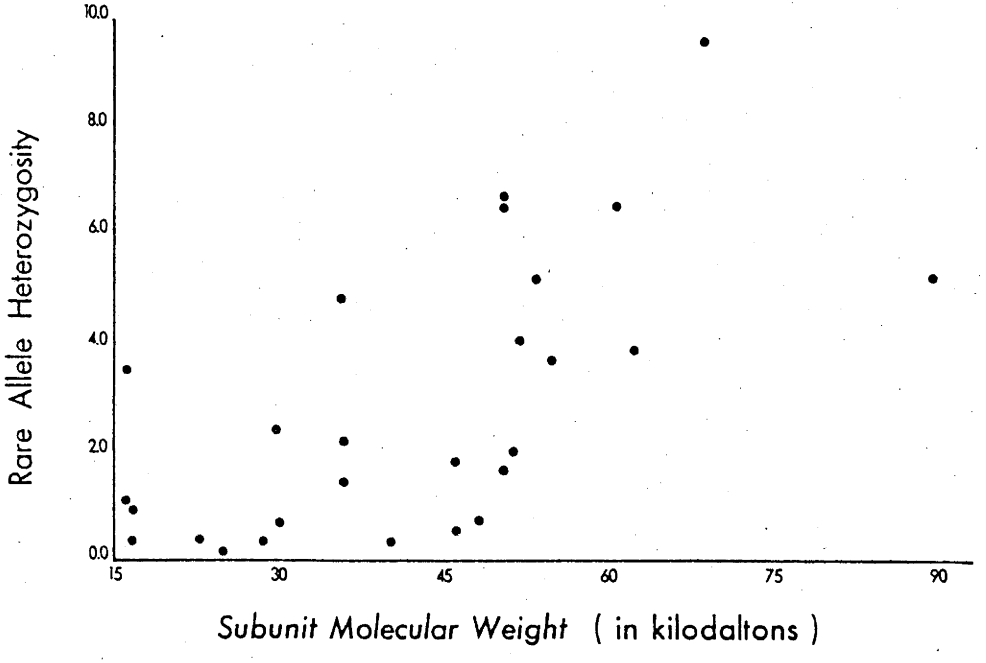

Fig. 5.2 Relationship between the rare allele heterozygosity (copies of rare alleles per 1,000 determinations) at a locus and the respective subunit molecular weights in human populations.

30.1 and 16.7 respectively. The most variant locus is albumin (Alb) for REMR(1) and phosphohexose isomerase (PHI) for REMR(2).

The ratios of REMR(1) and REMR(2) for multimers are 16.00 and 13.42 and for monomers, 30.14 and 10.38 respectively. In comparison the ratio of minimum to maximum subunit size is only 4.13 and 6.00 in multimers and monomers respectively. Thus the variability in REMRs is 3-5 times more than the observed variability in subunit size.

5.3.2. Interpopulational variability.

Since the data on various loci were compiled from different population groups, interpopulation comparisons have been made also. Table 5.4 shows the number of different rare alleles and rare allele heterozygosities in 12 human populations and the corresponding values of REMR(1) and REMR(2).

5.3.2.1. Number of rare alleles (K_r).

There is a large amount of variability in the detection of rare alleles among different populations. For example, in Iranians a new variant was detected for every 687 determinations whereas in non-Austronesians a rare allele was recovered for every 9,162 determinations. Although the detection of rare variants is a logarithmic function of sample size, it is interesting to note that there exists a negative correlation

TABLE 5.4 PARAMETERS OF RARE ALLELE VARIATION AND RELATIVE ELECTROMORPH MUTATION RATES IN 12 HUMAN POPULATIONS

Population	No. of determinations	Rare alleles		Rare allele heterozygosity (H_r)	No. of different rare alleles (K_r)	Relative electromorph mutation rates (REMR)	
		Number	Copies			REMR(1)	REMR(2)
	A	B	C	$D=\frac{C}{A} \times 1000$	$E=\frac{B}{A} \times 1000$	$F=\frac{D_i}{\Sigma D_i}$	$G=\frac{E_i}{\Sigma E_i}$
Amerindians	150,521	28	426	2.83	0.186	0.0907	0.0393
Japanese	70,388	37	175	2.49	0.526	0.0798	0.1112
English	109,009	42	83	0.76	0.385	0.0244	0.0814
Aust. Aborigines	66,258	14	170	2.57	0.211	0.0824	0.0446
Melanesians	263,890	37	992	3.76	0.140	0.1205	0.0296
Austronesians (AN)	89,806	17	163	1.82	0.189	0.0584	0.0399
Non-Austronesians (NAN)	174,084	19	556	3.19	0.109	0.1023	0.0230
S. Asian (Sch. Tribes)	65,974	20	178	1.66	0.303	0.0532	0.0640
S. African Negroes	39,865	15	46	1.15	0.376	0.0369	0.0078
S. African Khoisan	16,895	7	84	4.98	0.414	0.1596	0.0875
Aymaras	32,004	14	73	2.28	0.437	0.0731	0.0924
Iranians	10,993	16	41	3.72	1.455	0.1192	0.3075
Total (Except sample AN, NAN)	825,797	266	2,628	2.75	0.322	-	-

between the sample size and number of different rare alleles per 1,000 determinations ($r=-0.560$ $P<0.029$). At present it is difficult to give an explanation of this result.

As shown above, the recovery of rare alleles for the total population does not differ significantly between monomers and multimers. However, although the values are significant for the individual populations of Japanese, English, Australian Aborigines, S. Asian tribes, S. African Negroes and Aymaras, in the English and S. African Negroes, multimers show more rare variants; monomers are in excess in the other four (table 5.5).

5.3.2.2. Rare allele heterozygosity (H_r).

Significant heterogeneity in the interpopulation variability of rare allele heterozygosity (H_r) was detected over individual loci except GLO (no variant recovered), EsD and GOT. Coincidentally, the recovery of rare alleles at these loci is, in general, rather low. Chi-square heterogeneity over populations is also found to be significant for weighted mean values for multimeric, monomeric and all loci combined. The mean values of rare allele heterozygosity (H_r) in different populations range from 0.76 in English to 4.98 in South African Khoisans.

It has been pointed out previously that there is a significant difference in rare allele

TABLE 5.5 A COMPARISON OF RARE ALLELE HETEROZYGOSITY (H_r) AND NUMBER OF RARE ALLELES (K_r) BETWEEN MONOMERS AND MULTIMERS FOR 12 HUMAN POPULATIONS AND TOTAL SAMPLES

Population	Rare Allele Heterozygosity			Number of rare alleles per 1000 determinations	
	Multimers	Monomers	χ^2	Multimers	Monomers
Amerindians	2.53	3.19	5.76*	0.192	0.178
Japanese	1.30	4.20	57.57**	0.433	0.659
English	0.78	0.74	0.04	0.453	0.316
Australian Aborigines	1.99	3.27	10.54**	0.166	0.267
Melanesians	2.41	6.00	214.00**	0.152	0.121
Austronesians	1.95	1.58	1.52	0.193	0.187
Non-Austronesians	1.99	5.14	128.33**	0.130	0.076
S. Asian (Sch. Tribes)	2.49	1.15	1.80	0.244	0.401
S. African Negroes	1.83	0.46	16.26**	0.593	0.153
S. African Khoisan	7.46	1.96	25.61**	0.432	0.392
Aymaras	2.18	2.46	0.26	0.396	0.509
Iranians	4.36	2.68	1.95	1.598	1.217
Total sample	2.21	3.48	143.58**	0.249	0.238

* $0.01<P<0.05$

** $P<0.01$

heterozygosity between monomers and multimers for the total sample. The same effect is apparent when considering individual populations; significant differences being present between monomers and multimers for 7 out of 12 populations. In the two African populations, however, the rare allele heterzygosity is significantly higher in multimers. The reason for these differences is not clear.

5.3.2.3. Relative electromorph mutation rates (REMRs).

Table 5.4 shows the range of interpopulational estimates of REMR(1) and REMR(2). The range of REMR(1) is less than an order of magnitude; REMR(2) shows slightly more variation. The differences are smaller, in comparison with the effective population sizes of the groups in question.

No correspondence between monomers and multimers was detected for REMR(1) across 12 populations ($r=-0.036$; $P>0.485$). The results indicate lack of any systematic pressure on the frequency of mutation rates in different human populations, with regard to protein structure.

5.4. Discussion.

From the results outlined above it is apparent that in human populations sampled on an adequate scale, the size of molecules, and whether the intact molecule consists of a single subunit or of subunits combined together in multimers, has

an important influence on the relative magnitudes of electromorph mutation rates. The range of mean rare allele heterozygosity for these different categories is 3-4 times for both multimers and for all loci, when molecules with similar subunit molecular weights are compared. Monomers, on the other hand, reveal a larger variability, although the range is still less than ten-fold (see table 5.2).

Analysis of variance among categories reveals that this variability is real rather than stochastic ($F=6.96$; $P<0.01$) but the various parameters of rare allele variation, when normalized for subunit size, indicate non-significant variation among different categories. The molecules, in the middle range of subunit size, however, reveal least variability.

Some of the results given here are at variance with the previous analysis of data in the previous chapter on Australian Aborigines. Differences in mutation rate of more than an order of magnitude between multimers and monomers, after weighting for sample size, subunit size and adjusting for the proportion of aminoacids involved in molecular surface interactions in multimers, in the data on Australian Aborigines are not seen in the present data. The simplicity of proportionality between the molecular size and heterozygosity assumed in this and the previous chapter is, however, question-

able, especially when individual aminoacids, nucleotides and sites within cistrons are known to show variability in their substitution rates (Dayhoff *et al*, 1978; Kimura, 1979; Go and Miyazawa, 1980).

Although the magnitude of variability of the relative electromorph mutation rates estimated from rare allele heterozygosity and number of different rare alleles, is much smaller than that found by Zouros (1979) using total heterozygosity, the differences between the ranges of subunit size and REMRs are still significant. It appears that, since rare alleles are less likely to be operated upon by systematic negative or positive selection, comparatively large numbers of rare alleles may be maintained by slightly deleterious mutations (Ohta, 1976; Li, 1978, 1979a) or bottle-neck effect (Nei, 1976; Nei and Li, 1976). Although Bhatia (1980) and Chakraborty *et al*, (1980) found no correlation between the number of different rare alleles and total heterozygosity, the effect of intragenic recombination (Strobeck and Morgan, 1978; Morgan and Strobeck, 1979) on loci with unusually high mutation rates may be another factor contributing to this variability. The larger variability seen in the size of mRNAs (Sommer and Cohen, 1980) may also decide eventually the total amount of mutations obtained at a particular locus.

The range of interlocus variability in the estimates of H_r is, in general, much higher within populations than in the aggregate data. For example, Harris (1978) has recorded a 150 fold range in the values of H_r in English populations. The data presented here shows similar ranges in other major world populations. Genetic drift and geometric distributions of the copies of rare alleles (Rothman and Adams, 1978) are two of the reasons which can be invoked to explain this much larger variability. It may be relevant to point out here that only 6 out of 12 populations show significant correlations between the molecular weight and rare allele heterozygosity which indicates clearly that the relationship cannot be demonstrated unequivocally at the level of individual populations. Besides, recent fluctuations in population sizes may also affect these individual population correlations (Li, 1979b).

The distribution of REMRs does not give a good fit to either a gamma or lognormal distribution. Cavalli-Sforza and Bodmer (1971) and Yasuda (1973) have examined mutation rates using a lognormal and a gamma distribution respectively. Nei *et al.*, (1976b), Fuerst *et al.*, (1977), Chakraborty *et al.*, (1978), Zouros (1979) and Chakraborty *et al.*, (1980) have shown a gamma

distribution of mutation rates in proteins supported with similar evidence from distribution of protein subunit sizes. Sommer and Cohen (1980), however, found that the frequency distribution of subunit molecular weights, is well described by a lognormal distribution. While the subunit molecular weights of the loci included in the present study do, as shown by non-significant values of Pearson's statistics for their log values, follow lognormal distribution, the results for REMRs are not so well described by this distribution. One possibility is that compound distributions, which may arise from substituion processes at nucleotide level and distribution of cistron sizes, are involved.

It is interesting to consider if there exists any interpopulational correspondence in the single locus estimates of rare allele heterozygosity. Since the amount of normalized identiy (I) between any two distinct populations is negligible for rare alleles, the existence of such correlations must be a function of slightly deleterious mutations (Ohta, 1976) or variable mutation rates (Chakraborty *et al*, 1978). The significance of this correlation can be tested using normal tests since the value of r follows a normal distribution for I=0 (Chakraborty *et al*, 1978). The existence of such a relationship can be seen in the significant correlations between single locus rare allele heterozygosities of two samples

obtained from the same Japanese population by Neel *et al* (1980a). In the present study 13 of the 66 possible estimates of the coefficient of correlation for single locus rare allele heterozygosities among 12 populations are significant. Since half of the populations show significant correlations between rare allele heterozygosity and molecular weight, from the foregoing discussion it could be expected that 15 of the 66 pairwise comparisons will show significant correlations. The close approximation of the observed and expected number of significant correlations is encouraging.

Eanes and Koehn (1977), Chakraborty and Fuerst (1979) and Chakraborty *et al*, (1980) suggest that the correlation between the number of different alleles and molecular weights is generally higher than the correlation between molecular weight and heterozygosity. Chakraborty *et al*, (1980) confirm theoretically that this is expected to be so. For large sample sizes, they expect these correlations to be higher because of the inclusion of slightly deleterious mutations. For rare alleles, over sufficiently large sample sizes, this study shows the results to be otherwise for monomers and the total number of loci (table 5.3). The partial correlations, after controlling the sample size, however, confirm the observations of Eanes and Koehn (1977) and Chakraborty *et al*, (1980).

The magnitude of interpopulation variability in REMRs recorded in the present study is smaller than the interlocus variability. One of the factors influencing this is the amount of variability compressed within electromorphs which is related to $N\mu$ (Chakraborty and Nei, 1976; Nei and Chakraborty, 1976). Zouros (1979) has considered these differences to be the relative estimates of effective population sizes (N_e). However, demographic features of human populations have altered so much in the past and the errors involved in computing estimates of N_e are so large that I prefer to call these estimates interpopulational REMRs rather than relative effective population sizes (REPS).

The need for both absolute direct and indirect estimates of mutation rates in man from proteins has been emphasized by a number of workers (Neel, 1973, 1977; Neel and Rothman, 1978; Nei, 1977; Chakraborty and Roychoudhury, 1978; Dubinin and Altukhov, 1979; Tchen *et al.*, 1978; Bhatia *et al.*, 1979; Bhatia *et al.*, 1981; Bhatia, 1980). But it will be quite some time before reliable estimates are generated. With the accummulating evidence for the correlation of subunit size and molecular structure with mutation rates in animal and plant species, and now in man, estimates of relative electromorph mutation

rates can be extrapolated to real problems in population genetic theory.

Chapter 6

CONCLUSIONS

The estimates of mutation rate obtained in the present study correspond well, on an average within the order of magnitude, with earlier estimates generated by Neel (1973), Nei (1977), Neel and Rothman (1978), Neel and Thompson (1978), Tchen *et al.*, (1978), Chakraborty and Roychoudhry (1978) and Chakraborty (1981). The thirty eight individual populations for which the mutation rates have been generated in the present study, however, show a wide range in the estimates. While no result is obtained for 15 of the 38 populations (39.47%) because of the lack of recovery of private variants, the range for 23 non-null results of $\hat{\mu}_{K-0}$ and $\hat{\mu}_{R-A}$ is 0.95×10^{-6} - 33.42×10^{-6} per locus/generation and 0.65×10^{-6} - 35.33×10^{-6} per locus/generation respectively. The respective mean values are 5.99×10^{-6} and 6.39×10^{-6} per locus per generation (see table 6.1).

For 12 Amerindian populations on which the mutation rates were estimated by Neel and Rothman (1978), the respective values of $\hat{\mu}_{K-0}$ and $\hat{\mu}_{R-A}$ (with their ranges in parentheses) are 14.31×10^{-6} (0-36.87×10^{-6}) and 17.09×10^{-6} (0-44.05×10^{-6}) per locus per generation.

TABLE 6.1 DISTRIBUTION OF THE MUTATION RATE IN 38 HUMAN POPULATIONS

Mutation rate ($\times 10^{6}$)	Observed number of populations				
	μ K-O	μ R-A	μ NEI	μ CHAK	μ k_s
0.0 - 1.0	16	17	22	24	30
1.0 - 2.0	3	3	0	4	2
2.0 - 5.0	10	7	5	2	0
5.0 -10.0	6	4	4	2	2
10.0 - α	3	7	3	2	4
Total	38	38	34	34	38
Average ($\mu \times 10^{6}$)	5.99	6.39	4.62	2.55	2.86

The average estimates of μ obtained in the present study are less than half those obtained by Neel and Rothman (1978). This is because the proportion of null results in the present series is much higher (39.47%) than in that of Neel and Rothman (1978) where only 1 in 12 populations (8.33%) did not yield any result. The relative lack of recovery of private variants is probably due to the smaller sample sizes for the populations included in the present study than those utilized by Neel and Rothman.

The mumber of variants detected per locus decreases when more stringent criteria for defining an allelic variant, e.g. a rare variant or a singleton, are utilized. The average estimates of μ by the rare allele methods of Nei (1977) and Chakraborty (1981) and the singletons approach of the present study are thus much smaller than those obtained by the other two methods (table 6.1). These estimates also exhibit more variation in their inter-populational ranges, but this is reduced considerably with increase in the sample sizes.

The utility of the rare alleles and singletons methods is enhanced when the data from different populations are pooled because of the more assured recovery of rare or singleton variants. The estimates of mutation rate

obtained for five pooled samples of Australian Aborigines, Papua New Guineans, Scheduled Tribes of South India, Khoisans and Pygmies thereby exhibit much less variation (see table 6.2) although the mean estimates are still comparatively smaller than those obtained by the approaches of Kimura and Ohta (1969) and Rothman and Adams (1978).

One of the disadvantages in pooling the data in this form is that it completely disregards the underlying structure of populations and also lets the results be dominated by samples from a large population. According to Nei (1977) in the absence of deleterious genes, such a pooling should lead to estimates almost identical to those obtained through simple averaging. In the present study this is so (tables 6.1 and 6.2) except for μ_{NEI}. This particular problem arises because of extra-ordinarily high estimates generated for certain populations where $2nq<2.718$.

A comparison of the average of the results obtained in the present study with the weighted estimates (weighted by the population size) of $\hat{\mu}_{K-O}$, $\hat{\mu}_{NEI}$, $\hat{\mu}_{R-A}$ (Neel and Rothman, 1978) and μ_{CHAK} (Chakraborty 1981) as also the results on private polymorphisms (Neel and Thompson, 1978) indicates that the estimates of mutation rate are between $2x10^{-6}$ to $8x10^{-6}$ per locus/generation.

TABLE 6.2 ESTIMATES OF MUTATION RATES ($\mu x 10^6$)

IN VARIOUS POPULATIONS

Population	$\hat{\mu}x10^6$				
	$\hat{\mu}_{K-O}$	$\hat{\mu}_{R-A}$	$\hat{\mu}_{NEI}$	$\hat{\mu}_{CHAK}$	$\hat{\mu}_{k_s}$
Australian Aborigines	13.40	16.63	3.58	3.08	6.51
Papua New Guineans	2.83	6.58	1.44	1.30	3.80
Scheduled Tribes of India	0.73	2.44	0.26	0.23	0.36
Khoisans	4.80	1.47	1.23	1.14	0.77
Pygmies	4.11	1.26	1.02	0.85	-
Average	5.17	5.68	1.51	1.32	2.29

The above estimates of mutation rate are obtained on the basis of a single set of electrophoretic conditions only. While a number of charge-change variants are not detected through these electrophoretic screens because of the coalescence of a different number of variants into a single electromorph (Nei and Chakraborty, 1976; Chakraborty and Nei, 1976), only about one third of the aminoacid substitutions will actually lead to charge changes (Shaw, 1965; Nei and Chakraborty, 1973; Marshall and Brown, 1975). Adjusting for the silent substitutions of the latter type leads to threefold increase in the estimates of μ for the populations reported earlier. The averages over the five pooled samples range from 3.96×10^{-6} to 17.04×10^{-6} by five different methods. These estimates may, however, be considered conservative since there is still no adjustment made for the coalescence within electromorphs.

One of the most difficult aspects of these indirect estimates of mutation rate is the determination of the actual size of the population in question(N). This is particularly difficult in the case of the highly nomadic Australian Aborigines, Khoisan and Pygmies and the more continuously distributed, densely settled populations of South India and Papua New Guinea. In most of these populations no exact delimitation

of the isolate is possible. The estimates of the actual population size utilized above may thus, at best, be considered to be only close approximations.

Estimates of θ, which are not affected by the actual population size, for various populations studied here are given in table 6.3. These estimates show much smaller variation over populations than is noticed in the estimates of μ (see table 6.2). Since all the estimation procedures utilized here assume a stationary population over a large number of generations, such estimates of N are inaccurate in demographically retracting populations such as Australian Aborigines, Khoisans and Pygmies or in populations with positive growth trends like Scheduled Tribes of South India and Papua New Guineans. However, this is the only approach available at present.

The ratio $\hat{\theta}_{k_s} / \hat{\theta}_{k_t}$ obtained from the values given in table 6.3 shows the number of different alleles encompassed within an electromorph. Since the single copy electromorphs are composed of single alleles, the ratio is a good indicator of the electrophoretically silent alleles. This ratio exhibits a range of 1.38 to 4.40 alleles with an agerave of 2.80 alleles per electromorph, for an average value of $\hat{\theta}_{k_s} = 0.2043$. Nei and Chakraborty (1976) have provided this ratio

TABLE 6.3 COMPARATIVE ESTIMATES OF θ_k IN VARIOUS POPULATIONS

Population	Number of genes sampled (2n)	$\hat{\theta}_k$			$\hat{\theta}_{k_s} / \hat{\theta}_{k_t}$
		$\hat{\theta}_{k_t}$	$\hat{\theta}_{k_r}$	$\hat{\theta}_{k_s}$	
Australian Aborigines	5,214	0.0758	0.1150	0.2399	3.1649
Papua New Guineans	12,072	0.1192	0.1697	0.5239	4.3951
Scheduled Tribes of India	4,521	0.0549	0.1066	0.1669	3.0401
Khoisans	1,980	0.0655	0.1338	0.0908	1.3863
Pygmies	2,351	0.0494	0.0858	-	-
Average	5,228	0.0730	0.1222	0.2043	2,7986*

* $\Sigma\theta_{k_s} / \Sigma\theta_{k_t}$

between the number of alleles and electromorphs expected for 1,000 genes and θ=0.1 and 0.2 as 2.38and 3.15 respectively. The estimates of mean ratio, $\hat{\theta}_{k_s}/\hat{\theta}_{k_t}$ falls within the expected range, an interesting result when it is considered that only private alleles are used to estimate $\hat{\theta}_{k_t}$ and $\hat{\theta}_{k_s}$ and a large number of polymorphic alleles are excluded from the estimation of θ_{k_t}. In addition, the sample size has large impact on $\hat{\theta}_{k_s}$.

Correcting for this factor of under-estimation yields the estimates of total mutation rate by the methods of Kimura and Ohta (1969) and Rothman and Adams (1978) as 43.41×10^{-6} and 47.69×10^{-6} per locus per generation respectively (see table 6.4). No such correction is required when singletons are used. The correction factor for the rare variants accordingly becomes 1.6718 which gives the total mutation rate by the methods of Nei (1977) and Chakraborty (1981) as 7.57×10^{-6} and 6.92×10^{-6} respectively. It may be noticed that the rare alleles approaches and the singleton approach yield almost similar estimates of μ total.

Estimates of μ were also obtained by using the amount of expected heterozygosity in the five populations (table 6.5). Only those loci which are included in the study, have been used for

TABLE 6.4 ESTIMATES OF TOTAL MUTATION RATE BY VARIOUS ESTIMATION PROCEDURES

	Electro-morph mutation rates ($x10^6$)	Correction factor	Total mutation rate ($x10^6$)
Kimura and Ohta's method	5.17	3 x 2.7986	43.41
Rothman and Adams' method	5.68	3 x 2.7986	47.69
Nei's method	1.51	3 x 1.6718	7.57
Chakraborty's method	1.32	3 x 1.6718	6.92
Singletons method	2.29	3	6.87
Average	3.19		22.49

estimating the average heterozygosity ($h_j=1-\Sigma x_{ij}^2$). The individual estimates of $\hat{\theta}_F$ do not correspond well with the respective values of $\hat{\theta}_k$ given in table 6.3. However, the average value of θ_F, when corrected for an upward bias of 40% (Ewens and Gillespie, 1974) yields an estimate $\hat{\theta}_F^*$ of 0.0721. This average is almost similar to the value of 0.0730 of $\hat{\theta}_{k_t}$.

Although Bhatia (1980) and Chakraborty (1981) find the relation between the number of rare alleles and heterozygosity significant in Australian Aborigines and Amerindians respectively, no such correspondence is seen by Bhatia (1981) in 12 world populations. Similarly, no correlation between $\hat{\theta}_F$ and $\hat{\theta}_{k_t}$ is observed in these populations, although the correspondence in the average estimates of $\hat{\theta}_F^*$ and $\hat{\theta}_{k_t}$ is encouraging.

In chapters 4 and 5 I tried to bring out the role of subunit size and structure in changing the rate of mutation per cistron. The conclusion from these results is that although there exists a relationship between the number of rare alleles as also the rare allele heterozygosity and the molecular constraints, there is no simple regressing line available for the extrapolation of mutation rate. This is partly because the correlation between subunit size and mutation rate is incomplete and partly because the molecules with medium sizes attract, on an average, less

TABLE 6.5 ESTIMATES OF θ_F FOR VARIOUS POPULATIONS OBTAINED FROM THE PROTEIN LOCI INCLUDED IN THE STUDY

Population	Heterozygosity $(1-\Sigma x_i^2)$	θ_F	θ_F^*
Australian Aborigines	0.0417	0.0435	0.0311
Papua New Guineans	0.0748	0.0808	0.0577
Scheduled Tribes of India	0.0950	0.1050	0.0750
Khoisans	0.1067	0.1194	0.0853
Pygmies	0.1350	0.1561	0.1115
Average		0.1010	0.0721

$$\theta_F^* = 0.7143\theta_F$$

mutations. However, it is clear that the choice of results on single proteins for the mutation rates per cistron must be qualified by the size of the molecules.

The need for both direct and indirect estimates of mutation rate in man from protein data has been emphasized by a number of workers. Ten years ago, it would have been difficult to ask for more than finding that the order of magnitudes correspond (Cavalli-Sforza and Bodmer, 1971). The attemps to resolve this problem in more detail have revealed the existence of variation at cistron level greater than an order of magnitude. In the present state of the art two sets of estimates exist. By choosing to use the sampling formulations which do not incorporate the demographic features of the population, except in the estimation of N, one gets the total rate of mutation per cistron per generation as $\sim 7 \times 10^{-6}$. The estimated rate is 6-8 fold higher when the demographic features are taken into account. Any attempt to resolve these differences will go a long way toward obtaining more reliable estimates of mutation rate.

The question of whether certain populations have higher mutation rates (or lower selection against mutation at these loci) in certain populations (Neel, 1973; Neel *et al.*,1980c) seems to be unresolved by the present analysis.

BIBLIOGRAPHY

Adams, J.W. and Kasakoff, A.B.: Factors underlying endogamous group size, in *Population and Social Organization*, edited by Nag, M., The Hague, Mouton,pp. 147-174, 1975.

Administration of Papua New Guinea: *Report to the General Assembly of the United Nations for July 1970-June 1971.* Canberra, Australian Government Publishing Service, 1972.

Almeida, A. de: *Bushmen and Other non-Bantu Peoples of Angola.* Johannesburg, Witwatersrand University Press for Institute for the Study of Man in Africa, 1965.

Annual Reports: *Papua; New Guinea.* Canberra, Government Printer, 1951.

Arthur, E., Steel, C.M., Evans, H.J., Povey, S., Watson, B. and Harris, H.: Genetic studies on human lymphoblastoid cell lines. Isozymes and cytogenetic heterogeniety in a cell line with evidence for localization of Pep A locus in man. Ann. Hum. Genet. 39: 33-42, 1975.

Auerbach, C.: *Mutation Research.* London, Chapman and Hall, 1971.

Auerbach, C.: Forty years of mutation research: A pilgrim's progress. Heredity 40: 177-187, 1978.

Bannister, A. and Johnson, P.: *Namibia.* 1979.

Basu, A.: A demographic study of the Kota of Nilgiri Hills. J. Ind. Anthrop. Soc. 7: 29-45, 1972.

Basu, A.: Physical anthropological research in south India: a bibliographical review. J. Ind. Anthrop. Soc. 18: 187-213, 1978.

Basu, M.P.: A demographic profile of the Irula. Bull. Anthrop. Surv. India 16: 267-289, 1967.

Bateson, W.: A suggestion as to the nature of the "Walnut Comb in Fowls". Proc. Camb. Phil. Soc. 13: 1905, 1928 (c.f. Auerbach, 1971).

Beaumont, B., Nurse, G.T. and Jenkins, T.: Highland and lowland populations of Lesotho. Hum. Hered. 29: 42-49, 1979.

Beckenbach, A.T. and Prakash, S.: Examination of allelic variation at the hexokinase loci of *Drosophila pseudoobscura* and *Drosophila persimilis* by different methods. Genetics 87: 743-761, 1977.

Ben-Yosef, Y., Hungerford, M. and Nadler, H.L.: The nature of mutation in Krabbe disease. Am. J. Hum. Genet. 30: 644-652, 1978.

Beretta, M., Ranzani, G., Antonini, G., Neghi, M., Siccardi, A. and Santachiara-Benerecetti, A.S.: Genetic characterization of TWA Pygmies by the analysis of 15 erythrocytic enzymatic markers. Atti. Assoc. Genet. Ital. 22: 21-22, 1977.

Bernini, L.E., De Jong, W.W. and Meera Khan, P.: Varianti emglobiniche mella populazione tribale dell' Andhra Pradesh. Molteplicita del locus α^{Hb} nell' uomo. Atti. Assoc. Genet. Ital. 15: 191-194, 1970.

Bhatia, K.: Factors affecting electromprh mutation rates in man: an anlysis of data from Australian Aborigines. Ann. Hum. Biol. 7: 45-54, 1980.

Bhatia, K.: Rare allele heterozygosity and relative electromorph mutation rates in man. Ann. Hum. Biol. 8: 263-276, 1981a.

Bhatia, K.: The frequency of private electrophoretic variants and indirect estimates of mutation rate in Scheduled Tribes from South India. Actan 5: 67-87, 1981b.

Bhatia, K., Blake, N.M. and Kirk, R.L.: The frequency of private electrophoretic variants in Australian Aborigines and indirect estimates of mutation rate. Amer. J. Hum. Genet. 31: 731-740, 1979.

Bhatia, K., Blake, N.M., Serjeantson, S.W. and Kirk, R.L.: The frequency of private electrophoretic variants and indirect estimates of mutation rate in Papua New Guinea. Amer. J. Hum. Genet. 33: 112-122, 1981.

Birdsell, J.B.: Local group composition among Australian Aborigines: a critique of the evidence from fieldwork conducted since 1930. Curr. Anthropol. 11: 115-142, 1970.

Blake, N.M.: Genetic variation of red cell enzyme systems i Australian Aboriginal populations. Occasional Papers in Human Biology 2: 39-82, 1979.

Blake, N.M., Kirk, R.L., McDermid, E.M., Case, J. and Bashir, H.: The distribution of blood, serum protein and enzyme groups in a series of Lebanese in Australia. Aust. J. Exp. Biol. Med. Sci. 51: 209-220, 1973.

Blake, N.M., Kirk, R.L., McDermid, E.M., Omoto, K. and Ahuja, Y.R.: The distribution of serum protein and enzyme group systems among north Indians. Hum. Hered. 21: 440-457, 1971.

Blake, N.M. and Omoto, K.: Phosphoglucomutase types in the Asian-Pacific area: a critical review including new phenotypes. Ann. Hum. Genet., Lond. 38: 251-273, 1975.

Blake, N.M., Ramesh, A., Vijaykumar, M., Murty, J.B. and Bhatia, K.K.: Genetic studies on some tribes of the Telangana region, Andhra Pradesh. Actan 5: 41-56, 1981.

Bodmer, W.F. and Cavalii-Sforza, L.L.: Variation in fitness and molecular evolution. Proc. Sixth Berkeley Symp. Math. Stat. and Prob. 5: 255-275, 1972.

Bodmer, J.G. and Bodmer, W.F.: Studies on African Pygmies IV. A comparative study of the HL-A polymorphism in the Babinga Pygmies and other African and Caucasian populations. Am. J. Hum. Genet. 22: 396-411, 1970.

Botha, M.C., Toit, E.D. du, Jenkins, T., Leeuwen, A. van , d'Amaro, J., Meera Khan, P., Steen, G. van der , Rood, J.J. van and Does, J.A. van der : The HLA system in Bushman (San) and Hottentot (Khoikhoi) populations of South West Africa, in *Histocompatibility Testing 1972*, edited by Dausset, J. and Colombani, J., Copenhagen, Munksgaard, pp. 421-432, 1973.

Brock, D.J.H.: The structure and function of proteins, in *The Biochemical Genetics of Man*, edited by Brock, D.J.H. and Mayo, O., London, New York and San Francisco, Academic Press, 1978.

Brockfield, H.C. and White, J.P.: Revolution or evolution in the prehistory of the New Guinea highlands: a seminar report. Ethnology, 7: 43-52, 1968.

Brown, A.J.L. and Langley, C.H.: Correlations between heterozygosity and molecular weight. Nature 277: 649-651, 1979.

Brown, P. and Podolefsky, A.: Population density, agricultural intensity, land tenure and group size in the New Guinea Highlands. Ethnology 15: 211-238, 1976.

Bucher, K., Ionasescu, V. and Hanson, J.: Frequency of new mutants among boys with Duchenne muscular dystrophy. Am. J. Med. Genet. 7: 27-34, 1980.

Cavalli-Sforza, L.L.: Human diversity, in *Proc. Twelfth Int. Cong. Genet.*, Tokyo, Science Council of Japan 3: 405-416, 1969.

Cavalli-Sforza, L.L.: Pygmies as an example of hunter-gatherers and genetic consequences for man of domestication of plants and animals, in *Human Genetics*, edited by Grouchy, J. de., Ebling, F.J.G. and Henderson, I.W., Amsterdam, Excerpta Medica, pp. 79-95, 1972.

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

Cavalli-Sforza, L.L., Zonta, L.A., Nuzzo, F., Bernini, L., De Jong, W.W.W., Meera Khan, P., Ray, A.K., Went, L.N., Siniscalco, M., Nijenhuis, L.E., van Loghem, E. and Modiano, G.: Studies on African Pygmies .I. A pilot investigation of Babinga Pygmies in the Central African Republic (with an analysis of genetic distances). Am. J. Hum. Genet. 21: 252-274, 1969.

Census of India: *Special tables on Scheduled Castes and Scheduled Tribes. Series 2 Andhra Pradesh, Part V-A*. Hyderabad, Directorate Census Operations, 1971.

Chakraborty, R.A.: Estimation of mutation rates from the number of rare alleles in a sample. Ann. Hum. Biol. 8: 221-230, 1981.

Chakraborty, R. and Fuerst, P.A.: Some sampling properties of selectively neutral alleles. Genet. Res. 34: 253-267, 1979.

Chakraborty, R., Fuerst, P.A. and Nei, M.: Statisical studies on protein polymorphism in natural populations II. Gene differentiation between populations. Genetics 88: 367-390, 1978.

Chakraborty, R., Fuerst, P.A. and Nei, M.: Statistical studies on protein polymorphism in natural populations III. Distribution of allele frequencies within populations. Genetics 94: 1039-1063, 1980.

Chakraborty, R. and Nei, M.: Hidden genetic variability within electromorphs in finite populations. Genetics 84: 385-393, 1976.

Chakraborty, R. and Roychoudhury, A.K.: Mutation rates from rare variants of proteins in Indian tribes. Hum. Genet. 43: 179-183, 1978.

Chen, S.H., Giblett, E.R., Anderson, J.E. and Fossum, B.L.G.: Genetics of glutamic pyruvic transaminase: its inheritance, common and rare variants, population distribution and differences in catalytic activity. Ann. Hum. Genet., Lond. 35: 401-409, 1972.

Classen, O.R. and McElhanon, K.A.: Languages of the Finnisterre Range - New Guinea, in *Papers in New Guinea Linguistics, Ser. A., Occasional Papers No. 23, Pacific Linguistics,* Canberra, The Australian National University, pp. 45-78, 1970.

Cleaver, J.E.: Defective DNA repair and hereditary diseases. Abs. Inter. Conf. on Defective DNA Repair, Mutation and Human Ill Health. Ottawa, May 8-10, 1978.

Coale, A.J. and Demeny, P.: *Regional Model Life Tables and Stable Populations.* Princeton, Princeton Univ. Press, 1966.

Commonwealth of Australia: *Population and Australia.* Canberra, Government Printing Service, 1975.

Conneally, P.M.: Mutation rates in Man, in *Modern Trends in Human Genetics,* edited by Emery, A.E.H. London and Boston, Butterworths 2: 204-220, 1974.

Coon, C.S.: *The Living Races of Man.* New York, Knopf, 1965.

Coyne, J.A. and Felton, A.A.: Genic heterogeneity at two alcohol dehydrogenase loci in *Drosophila pseudoobscura* and *Drosophila persimilis.* Genetics 87: 285-304, 1977.

Coyne, J.A., Felton, A.A. and Lewontin, R.C.: Extent of genetic variation at a highly polymorphic esterase locus in *Drosophila pseudoobscura.* Proc. Natl. Acad. Sci. USA 75: 5090-5093, 1978.

Crow, J.F.: Some possibilities for measuring selection intensities in man. Hum. Biol. 30: 1-13, 1958.

Crow, J.F., Mutation in man. Prog. Med. Genet. 1: 1-26, 1961.

Crow, J.F. and Kimura, M.: The effective number of a population with overlapping generations: a correction and further discussion. Am. J. Hum. Genet. 24: 1-10, 1972.

Crow, J.F. and Morton, N.E.: Measurement of gene frequency drift in small populations. Evolution 9: 202-214, 1955.

Danforth, G.H.: The frequency of mutations and the incidence of hereditary traits in man. *Eugenics, Genetics and Family*. Scientific papers of the 2nd Int. Cong. Eugenics, New York, 1: 120-128, 1921.

Darnall, D.W. and Klotz, I.M.: Subunit constitution ofproteins: A table. Arch. Biochem. Biophys. 166, 651-682, 1975.

Davie, A.M. and Emery, A.E.H.: Estimation of proportion of new mutants among cases of Duchenne muscular dystrophy. J. Med. Genet. 15: 339-345, 1978.

Dayhoff, M.O., Schwartz, R.M. and Orcutt, B.C.: A model of evolutionary change in proteins, in *Atlas of protein sequence and structure*, edited by Dayhoff, M.O., Washington, National Biomedical Research Foundation, Vol 5, Supp. 3. pp. 345-352, 1978.

De Jong, W.W.W., Bernini, L.E. and Meera Khan, P.: Haemoglobin Rampa: α95 Pro→Ser. Biochim. Biophys. Acta 236: 197-200, 1971.

De Jong, W.W.W., Meera Khan, P. and Bernini, L.E.: Haemoglobin Koya Dora: high frequency of a chain termination mutant. Amer. J. Hum. Genet. 27: 81-90, 1975.

Department of Chief Minister and Development Administration: *Unpublished records*, Waigani, Papua New Guinea, 1970-74.

Department of Foreign Affairs, RSA: *South West Africa Survey, 1967*. Pretoria and Capetown, Government Printer, 1967.

Dewey, W.J., Barrai, I., Morton, N.E. and Mi, M.P.: Recessive genes in severe mental defect. Am. J. Hum. Genet. 17: 237-256, 1965.

Drake, J.W.: Basic mutational mechanisms and human health. Abs. Inter. Conf. on Defective DNA Repair, Mutation and Human I 11 Health. Ottawa. May 8-10, 1978.

Dubinin, N.P. and Altukhov, Yu. P.: Gene mutations (*de novo*) found in electrophoretic studies of blood proteins of infants with anomalous development. Proc. Natl. Acad. Sci. USA. 76: 5226-5229, 1979.

Dutton, H.E.: *Languages of the South-East Papua: A Preliminary Report. Ser. A. No. 28, Pacific Linguistics*. Canberra, The Australian National University, pp. 1-46, 1971.

Eanes, W.F. and Koehn, R.K.: The correlation of rare alleles with heterozygosity: determination of the correlation for neutral models. Genet. Res. 29: 223-230, 1977.

Eanes, W.F. and Koehn, R.K.: Relationship between subunit size and number of rare electrophoretic alleles in human enzymes. Biochem. Genet. 16: 971-985, 1978.

East African Statistical Department: *Census 1957. General African Census, August 1957, Part I, Tribal Analysis* · Nairobi, 1958.

Edwards, J.H.: The mutation rate in man. Prog. Med. Genet. 10: 1-16, 1974.

Elston, R.C.: Ascertainment and age of onset in pedigree analysis. Hum. Hered. 23: 105-112, 1973.

Elston, R.C. and Stewart, J.: A general model for the genetic analyis of pedigree data. Hum. Hered. 21: 523-542, 1971.

Elston, R.C. and Yelverton, K.C.: General models for segregation analysis. Am. J. Hum. Genet. 27: 31-45, 1975.

Esterman, C.: *Etnographia do Sudoeste de Angola Vol. I. Os Povos Não-Bantos e o Grupo Étnico dos Ambós*. Porto, Ministerio do Ultramar, 1956.

Ewens, W.J.: The maintenance of alleles by mutation. Genetics 50: 891-898, 1964.

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

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

Ewens, W.J.: Population genetics theory in relation to the neutralist-selectionist controversy. Adv. Hum. Genet. 8: 67-133, 1977.

Ewens, W.J.: Testing the generalized neutrality hypothesis. Theor. Pop. Biol. 15: 205-216, 1979a.

Ewens, W.J.: *Mathematical Population Genetics.* Berlin, Heidelburg and New York, Spring-Verlag, 1979b.

Ewens, W.J. and Feldman, M.W.: The theoretical assessment of selective neutrality, in *Population Genetics and Ecology*, edited by Karlin, S. and Nevo, E., New York, Academic Press, pp. 303-317, 1976.

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

Ewens, W.J. and Li, W-H.: Frequency spectra of neutral and deleterious alleles in a finite population. J. Math. Biol. 10: 155-166, 1980.

Feller, W.: *An Introduction to the Probability theory and its applications.* New York, John Wiley, 1950.

Fisher, R.A.: *The genetical theory of natural selection.* Oxford, Clarendon Press, 1930.

Francke, U., Felsenstein, J., Gartler, S.M., Migeon, B.R., Dancis, J., Seegmiller, J.E., Barkay, F. and Nyhan, W.L.: The occurrence of new mutants in the X-linked recessive Lesch-Nyhan disease. Am. J. Hum. Genet. 28: 123-137, 1976.

Francke, U., Felsenstein, J., Gartler, S.M., Nyhan, W.L. and Seegmiller, J.E.: Answer to criticism of Morton and Lalouel (letter to the editor). Am. J. Hum. Genet. 29: 307-311, 1977.

Freese, E.: The specific mutagenic effects of base analogues onPhage T4. J. Mol. Biol. 1: 87-105, 1959.

Fuerst, P.A., Chakraborty, R. and Nei, M.: Statistical studies on protein polymorphism in natural populations. I. Distribution of single locus heterozygosity. Genetics 86; 455-483, 1977.

Gajdusek, D.C. and Alpers, M.: Genetic studies in relation to kuru I. Cultural, historical and demographic background. Am. J. Hum. Genet. 24: S1-S38, 1972.

Gardner-Medwin, D.: Mutation rate in Duchenne type of muscular dystrophy. J. Med. Genet. 7: 334-337, 1970.

Ghosh, A.K.: Selection intensity in Kota of Nilgiri Hills, Madras. Soc. Biol. 17: 224-225, 1970.

Ghosh, A.K.: Inbreeding in the Kota of Nilgiri Hills, Madras. Soc. Biol. 19: 289-291, 1972.

Ghosh, A.K.: The Kota of the Nilgiri Hills: a demographic study. J. Biosoc. Sci. 8: 17-26, 1976.

Ghosh, A.K.: The distribution of genetic variants of glyoxalase I, esterase D and carbonic anhydrase I and II in Indian populations. Indian J. Phys. Anthrop. Hum. Genet. 3: 73-83, 1977a.

Ghosh, A.K.: Polymorphism of red cell glyoxalase I with reference to south-east Asia and Oceania. Hum. Genet. 39: 91-95, 1977b.

Ghosh, A.K., Kirk, R.L., Joshi, S.R. and Bhatia, H. M.: A population genetic study of the Kota in the Nilgiri Hills, south India. Hum. Hered. 27: 225-241, 1977.

Gierer, A. and Mundry, K.W.: Production of mutants of tobacco mosaic virus by chemical alteration of its ribonucleicacid *in vitro*. Nature 182; 1457-1458, 1958.

Gillespie, J.H. and Kojima, K.: The degree of polymorphisms in enzymes involved in energy production compared to that in nonspecific enzymes in two *Drosophila ananassae* populations. Proc. Natl. Acad. Sci. USA 61: 582-585, 1968.

Go, M. and Miyazawa, S.: Relationship between mutability, polarity and exteriority of amino-acid residues in protein evolution. Inter. J. Pep. Res. 15, 211-224, 1980.

Godber, M., Kopec, A.C., Mourant, A.E., Teesdale, P., Tills, D., Weiner, J.S., El-Niel, H., Wood, C.H. and Barley, S.: The blood groups, serum groups, red cell isoenzymes and haemoglobins of the Sandawe and Nyature of Tanzania. Ann. Hum. Biol. 3: 463-473, 1976.

Goud, J.D. and Rao, P.R.: Distribution of some genetic markers in the Yerukala tribe of Andhra Pradesh. J. Ind. Anthrop. Soc. 12: 258-265, 1977.

Goud, J.D. and Rao, P.R.: Transferrin, haptoglobin and group specific component types in tribal populations of Andhra Pradesh. Hum. Hered. 30: 12-17, 1980.

Guerreiro, M.V.: *Bochimanes !Khū de Angola.* Lisboa, Junta de Investigacões do Ultramar, 1968.

Gunther, M. and Penrose, L.S.: The genetics of epiloia. J. Genet. 31: 413-430, 1935.

Gusinde, M.: *Von gelben und Schwarzen Buschmännern.* Graz, Akademische Druck-und Verlagsanstalt, 1966.

Haldane, J.B.S.: The rate of spontaneous mutation of a human gene. J. Genet. 31: 317-326, 1935.

Haldane, J.B.S.: The mutation rate of the gene for haemophilia and its segregation ratio in males and females. Ann. Eugen. (Lond) 13: 262-271, 1947.

Harris, H.: Multiple allelism and isozyme diversity in human populations, in *Mutations, Biology and Society*, edited by Walcher, D.N., Ketchner, N. and Barnett, H.C., New York, Mason Publishing House, pp. 77-98, 1978.

Harris, H.: *Principles of Human Biochemical Genetics.* Amsterdam and Oxford, North Holland Publishing Co., 1981.

Harris, H., Hopkinson, D.A. and Edwards, Y.H.: Polymorphism and the subunit structure of enzymes: A contribution to the neutralist-selectionist controversy. Proc. Natl. Acad. Sci. USA 74: 698-701, 1977.

Harris, H., Hopkinson, D.A. and Robson, E.B.: The incidence of rare alleles determining electrophoretic variants. Data on 43 enzyme loci in man. Ann. Hum. Genet. 37: 237-253, 1974.

Harpending, H.: Regional variation in !Kung populations, in *Kalahari Hunter-gatheres*, edited by Lee, R.B. and De Vore, I., Cambridge, Massachusetts and London, Harvard Univ. Press, pp. 152-165, 1976.

Harpending, H. and Jenkins, T.: !Kung population structure, in *Genetic Distance*, compiled by Crow, J.F. and Denniston, C., New York and London, Plenum Press, pp. 137-159, 1971.

Healey, A.: Austronesian Languages: Admiralty Islands area, in *New Guinea Area Languages and Language Study, Vol II* edited by Wurm, S.A., Ser.C., No. 38, Pacific Linguistics, Canberra, The Australian National University, pp. 349-364, 1975.

Hethcote, H.W. and Knudson, A.G.: Model for incidence of embryonal cancers: Application to retinoblastoma. Proc. Natl. Acad. Sci. USA 75: 2453-2457, 1978.

Hiernaux, J.: Physical Anthropology of the living populations of sub-Saharan Africa. Ann. Rev. Anthropol. 6: 149-168, 1976.

Hitzeroth, H.W. and Hummel, K.: Serum protein polymorphisms Hp, Tf, Gc, Gm, Inv and Pt in Bantu-speaking South African Negroids. Anthrop. Anz. 36: 127-141, 1978.

Hitzeroth, H.W., Walter, H., Hilling, M. and Munderloh, W.: Genetic markers and leprosy in South African Negroes. Part II. Erythrocyte enzyme polymorphisms. S.A. Med. J. 56: 507-510, 1979.

Hopkinson, D.A., Coppock, J.S., Muhlemann, M.F. and Edwards, Y.H.: The detection and differentiation of the products of the human carbonic anhydrase loci, CAI and CAII, using fluorogenic substrates. Ann. Hum. Genet., Lond. 38: 155-162, 1974.

Hopkinson, D.A., Edwards, Y.H. and Harris, H.: The distribution of subunit numbers and subunit sizes of enzymes: A study of the products of 100 human gene loci. Ann. Hum. Genet. 39: 383-411, 1976.

Hopkinson, D.A., Mestringer, M.A., Cortner, J. and Harris, H.: Esterase D: a new human polymorphism. Ann. Hum. Genet., Lond. 37: 119-137, 1973.

Hornabrook, R.W.: The demography of the population of Karkar Island. Phil. Trans Roy. Soc. Lond. B268 : 229-239, 1974.

Howell, N.: The population of Dobe Area !Kung, in *Kalahari Hunter-Gatherers*, edited by Lee, R.B. and De Vore, I., Cambridge, Massachusetts and London, Harvard Univ. Press, pp. 137-151, 1976.

Howell, N.: *Demography of the Dobe !Kung*. New York, San Francisco and London, Academic Press, 1979.

Ishimoto, G. and Kuwata, M.: Red cell glutamic-pyruvic transaminase polymorphism in Japanese populations. Jap. J. Hum. Genet. 18: 373-377, 1974.

Jenkins, T.: Blood group ^{A} *bantu* populations and family studies. Vox. Sang. 26: 537-550, 1974.

Jenkings, T., Harpending, H.C., Gordon, H., Keraan, M.M. and Johnston, S.: Red cell enzyme polymorphisms in Khoisan peoples of Southern Africa. Am. J. Hum. Genet. 23: 513-532, 1971.

Jenkins, T., Harpending, H.C. and Nurse, G.T.: Genetic distances among certain Southern African populations, in *Evolutionary Models and Studies in Human Diversity*, edited by Meier, R.J., Otten, C.M. and Abdel-Hamee, F., The Hague and Paris, Mouton Publishers, pp. 227-243, 1978.

Jenkins, T., Lane, A.B., Nurse, G.T. and Tanaka, J.: Sero-genetic studies on the G/wi and G//ana of Botswana. Hum. Hered. 25: 318-328, 1975.

Jenkins, T. and Nurse, G.T.: Biomedical studies on the desert-dwelling hunter-gatherers of Southern Africa, in *Progress in Medical Genetics, n.s., Vol. I*, edited by Steinberg, A.G., Bearn, A.G., Motulsky, A.G. and Childs, B., pp. 211-281, 1976.

Jenkins, T. and Steinberg, A.G.: Some serum protein polymorphisms in Kalahari Bushmen and Bantu: gammaglobulins, haptoglobins and transferrins. Am. J. Hum. Genet. 18: 399-407, 1966.

Johnson, G.B.: Enzyme polymorphism and metabolism. Science 184: 28-37, 1974.

Johnson, G.B.: Hidden heterogeneity among electrophoretic alleles, in *Measuring Selection in Natural Populations*, edited by Chrstiansen, F.B.

and Fenchel, T.M., Berlin, Springer-Verlag, pp. 223-244, 1977a.

Johnson, G.: Isozymes, allozymes and enzyme polymorphism: structural constraints on polymorphic variation, in *Isozymes II. Current Topics in Biological and Medical Research*, edited by Rattazi, M.C., Scandalios, J.G. and Whitt, G.S., New York, Allan R. Liss, Inc., pp. 11-19, 1977b.

Johnson, G.: Genetically controlled variation in the shapes of enzymes. Prog. Nucleic Acid Res. Mol. Biol. 22: 293-326, 1979.

Juke, T.H.: The amino acid code. Adv. Enzym. 47: 375-432, 1978.

Karlin, S. and McGregor, J.: The number of mutant forms maintained in a population. Proc. Fifth Berk. Sym. Math. Stat and Prob. 4: 415-438, 1967.

Karlin, S. and McGregor, J.: Addendum to a paper of W. Ewens. Theor. Pop. Biol. 3: 113-116, 1972.

Kazazian, H.H., Cho, S. and Phillips, J.A.: The mutational basis of the Thalassemia syndromes. Prog. Med. Genet. n.s. 2: 165-204, 1977.

Kimura, M.: Diffusion models in population genetics. J. Appl. Prob. 1: 177-232, 1964.

Kimura, M.: Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. Genet. Res. 11: 247-269, 1968.

Kimura, M.: The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics 61: 893-903, 1969.

Kimura, M.: The neutral theory of molecular evolution and polymorphism. Scientia 112: 687-707, 1977.

Kimura, M.: The neutral theory of molecular evolution. Sci. Amer. 241: 94-105, 1979a.

Kimura, M.: Model of effectively neutral mutations in which selective constraint is incorporated. Proc. Natl. Acad. Sci. USA 76: 3440-3444, 1979b.

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

Kimura, M. and Maruyama, T.: Pattern of neutral polymorphism in a geographically structured population. Genet. Res. 18: 125-131, 1971.

Kimura, M. and Ohta, T.: The average number of generations until extinction of an individual mutant gene in a finite population. Genetics 63: 701-709, 1969.

Kimura, M. and Ohta, T.: *Theoretical Aspects of Population Genetics*. Princeton, N.Y., Princeton Univ. Press, 1971.

Kimura, M. and Ohta, T.: Mutation and evolution at molecular level. Genetics 73: s19-s35, 1973.

Kimura, M. and Ohta, T.: On some principles governing molecular evolution. Proc. Natl. Acad. Sci. USA 71: 2848-2852, 1974.

Kimura, M. and Ohta, T.: Distribution of allele frequencies in a finite population under stepwise production of neutral alleles. Proc. Natl. Acad. Sci. USA 72: 2761-2764, 1975.

King, J.L. and Jukes, T.H.: Non-Darwinian evolution: Random fixation of selectively neutral mutations. Science 164: 788-798, 1969.

Kirk, R.L.: *Man Adapting: Human Biology of Australian Aborigines*. Oxford, Oxford University Press, 1981.

Kirk, R.L., Cleve, H. and Bearn, A.G.: The distribution of the group specific component (Gc) in selected populations in South and S.E. Asia and Oceania. Acta Genet., Basel 13: 140-149, 1963.

Kirk, R.L., Lai, L.Y.C., Vos, G.H., Wickremsinghe, R.L. and Perera, D.J.B.: The blood and serum groups of selected populations in south India and Ceylon. Amer. J. Phys. Anthrop. 20: 485-497, 1962.

Knudson, A.G.: Mutation and cancer: Statistical study of retinoblastoma. Proc. Natl. Acad. Sci. USA 68: 820-823, 1971.

Knudson, A.G.: Our load of mutations and its burden of disease. Amer. J. Hum. Genet. 31: 401-413, 1979.

Knussmann, R.: Bericht über eine anthropologische Forschungsriese den Dama in Sudwestafrika. Homo 20: 34-66, 1969.

Knussmann, R. and Knussmann, R.: Die Dama - eine Altschicht in Sudwestafrika? Jour. South West Afr. miss. Ges. 24: 9-32, 1969/70.

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

Koehn, R.K. and Eanes, W.F.: Subunit size and genetic variation of enzymes in natural populations of *Drosophila.* Theor. Pop. Biol. 11: 330-341, 1977.

Koehn, R.K. and Eanes, W.F.: Molecular structure and protein variation within and among populations. Evol. Biol. 10: 39-100, 1978.

Kömpf, J., Bissbort, S., Gussmann, S. and Ritter, H.: Polymorphism of red cell glyoxalase I (E.C.: 4.4.1.5.): a new genetic marker in man. Humangenetik 27: 141-143, 1975.

Lange, K., Gladstein, K. and Zatz, M.: Effect of reproductive compensation and genetic drift on X-linked lethals. Am. J. Hum. Genet. 30: 180-189, 1978.

Laycock, D.C.: The Torricelli phylum, in *New Guinea Area and Languages Study*, Vol. I, edited by Wurm, S.A., Ser. C, No. 38, Pacific Linguistics, Canberra, The Australian National University, pp. 767-780, 1975.

Laycock, D.C. and Z'Graggen, J.: The Sepik Ramu phylum, in *New Guinea Area Languages and Language Study,* Vol. I, edited by Wurm, S.A., Ser. C, No. 38, Pacific Linguistics, Canberra, The Australian National University, pp. 731-763, 1975.

Lea, D.E.: *Actions of Radiations on Living Cells.* New York, Cambridge Univ. Press, 1946.

Lea, D.E. and Coulson, C.A.: The distribution of the number of mutants in bacterial populations. J. Genet. 49: 264-285, 1949.

Lee, R.B.: Introduction, in *Kalahari Hunter-Gatherers,* edited by Lee, R.B. and De Vore, I., Cambridge, Massachusetts and London, Harvard University Press, pp. 3-24, 1976.

Lee, R.B.: *The !Kung San.* Cambridge, The Cambridge University Press, 1979.

Lee, R.B. and De Vore, I (eds): *Man the Hunter, Introduction.* Chicago, Aldine, 1968.

Lepervenche, M.: Social structure, in *Encyclopedia of Papua and New Guinea,* edited by Ryan, D., Melbourne, Melbourne Univ. Press, 1972.

Lewis, W.H.P. and Harris, H.: Peptidase D (prolidase) variants in man. Ann. Hum. Genet. 32: 317-322, 1969.

Li, F.H.F. and Neel, J.V.: A simulation of the fate of a mutant gene of neutral selective value, in *Computer Simulation in Human Populations,* edited by Dyke, B. and MacCluer, J.W., New York, Academic Press, 1974.

Li, F.H.F., Neel, J.V. and Rothman, E.D.: A second study of the survival of a neutral mutant in a simulated Amerindian population. Am. Nat. 112: 83-96, 1978.

Li, W-H.: Maintenance of genetic variability under mutation and selection pressures in a finite population. Proc. Natl. Acad. Sci. USA 74: 2509-2513, 1977.

Li, W-H.: Maintenance of genetic variability under the point effects of mutation selection and random drift. Genetics 90: 349-382, 1978.

Li, W-H.: Maintenance of genetic variability under the pressure of neutral and deleterious mutations in a finite population. Genetics 92: 647-667, 1979a.

Li, W-H.: Effect of changes in population size on the correlation between mutation rate and heterozygosity. J. Mol. Evol. 12: 319-329, 1979b.

Lieberman, G.J. and Owen, D.B.: *Tables of the hypergeometric probability distribution.* Stanford, Stanford Univ. Press, 1961.

Magni, G.E. and von Borstel, R.C.: Different rates of spontaneous mutation during mitosis and meiosis in yeast. Genetics 47: 1097-1108, 1962.

Majumdar, P.P.: Matrimonial migration: a review with special reference to India. J. Biosoc. Sci. 9: 381-401, 1977.

Marais, J.S.: *The Cape Coloured People, 1652-1937*. London, Longmans, Green and Co., 1939.

Marks, S.: 'Khoisan resistance to the Dutch in the Seventeenth and Eighteenth Centuries. J. Afr. Hist. 13: 55-80, 1972.

Marshall, D.R. and Brown, A.H.D.: The charge state model of protein polymorphism in natural populations. J. Mol. Evol. 6: 149-163, 1975.

Marshall, L.: *The !Kung of Nyae Nyae*. ,Cambridge, Massechusetts and London, Harvard Univ. Press, 1976.

Masters, C.J. and Holmes, R.S.: *Hemoglobin, Isoenzymes and Tissue Differentiation*. Amsterdam and Oxford, North Holland Publishing Co., 1975.

Matsunaga, E.: Hereditary retinoblastoma: Delayed mutation or host resistance. Am. J. Hum. Genet. 30: 406-424, 1978.

McCommas, S.A. and Chakraborty, R.: Estimation of mutation rates in three species of sea anemone in the genus *Bundosoma*. Genetics 94: s66, 1980.

McDermid, E.M. and Vos, G.H.: Serum protein groups of South African Bantu II. α_1-antitrypsin, group specific component and further observations of haptoglobin and caeruloplasmin. S. Afr. J. Med. Sci. 36: 63-68, 1971.

McElhanon, K.A.: The north-eastern areas of the trans New Guinea phylum: north-eastern trans-New Guinea phylum languages, in *New Guinea Area Languages and Language Study*, vol. I, edited by Wurm, S.A., Ser. C., No. 38, Pacific Linguistics, Canberra, The Australian National University, pp. 527-568, 1975.

Meggitt, M.J.: *Desert People*. Sydney, Australia, Angus and Robertson, 1962.

Middleton, M.R. and Francis, S.H.: *Yuendumu and its children*. Canberra, Australian Government Publishing Service, 1976.

Milliken, E.P.: Aboriginal language distribution in Northern Territory, in *Tribes and Boundaries in Australia*, edited by Peterson, N., Canberra, Australian Institute of Aboriginal Studies, 1976.

Moodie, D.: *The Record or A Series of Official Papers Related to the Condition and Treatment of the Native Tribes of South Africa.* Amsterdam, Balkema, 1940-42.

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

Morton, N.E.: The genetic tests under incomplete ascertainment. Am. J. Hum. Genet. 11: 1-16, 1959.

Morton, N.E.: The mutational load due to determinental genes in man. Am. J. Hum. Genet. 12: 348-364, 1960.

Morton, N.E.: Isolation by distance in human population. Ann . Hum. Genet., 40: 361-365, 1977.

Morton, N.E.: Genetic epidemiology in pedigrees: Kinship and path analysis. Rev. Brasil. Genet. II, 1: 1-15, 1979.

Morton, N.E., Crow, J.F. and Muller, H.J.: An estimate of the mutation damage in man from data on consanguineous marriages. Proc. Natl. Acad.Sci. USA 42: 855-863, 1956.

Morton, N.E., Klein, D., Mussels, I.E., Dodinval, P., Todorov, A., Lew, R. and Yee, S.: Genetic structure of Switzerland. Am. J. Hum. Genet. 25: 347-361, 1973.

Morton, N.E. and Lalouel, J.M.: Genetic counselling in sex linkage. Birth defects: Orig. Art. Ser. 15 (5c): 9-24, 1979.

Morton, N.E., Rao, D.C., Lang-Brown, H., MacLean, C.J., Bart, R.D. and Lew, R.: Colchester revisited: A genetic study of mental defect. J. Med. Genet. 14: 1-9, 1977.

Mukai, T.: Spontaneous mutation rates of isozyme genes in *Drosophila melanogaster.* Dros. Info. Surv. 45: 99, 1970.

Mukai, T.: Polygenic mutation, in *Quantitative Genetic Variation,* edited by Thompson, J.N., Jr. and Thoday, J.N., New York, Academic Press, pp. 177-196, 1979.

Mukai, T. and Cockerham, C.C.: Spontaneous mutation rate of enzyme loci in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA 74: 2514-2517, 1977.

Muller, H.: Variation due to change in the individual gene. Am. Nat. 56: 32-50,1922.

Murdock, G.P.: *Africa, Its People and Their Culture History*. New York, McGraw-Hill, 1959.

Murty, J.S. and Ramesh, A.: Selection intensities among the tribal populations of Adilabad District, Andhra Pradesh, India. Soc. Biol. 25: 302-305, 1979.

Neel, J.V.: Mutations in the human populations, in *Methodology in Human Genetics*, edited by Burdette, W.J., San Francisco, Holden-Day, Inc., 1962.

Neel, J.V.: "Private" genetic variants and the frequency of mutation among South American Indians. Proc. Natl. Acad. Sci. USA 70: 3311-3315, 1973.

Neel, J.V.: Some trends in the study of spontaneous and induced mutation in man. In *Human Genetics*, edited by Armendares and Lisker, R. Amsterdam, Excerpta Medica, pp. 19-32, 1977.

Neel, J.V.: Mutation and disease in man. Can. J. Genet. Cytol. 20: 295-306, 1978a.

Neel, J.V.: Rare variants, private polymorphisms and locus heterozygosity in Amerindian populations. Amer. J. Hum. Genet. 30: 465-470, 1978b.

Neel, J.V., Mohrenweiser, H.W. and Meisler, M.H.: Rate of spontaneous mutation at human loci encoding protein structure. Proc. Natl. Acad. Sci. USA 77: 6037-6041, 1980a.

Neel, J.V., Satoh, C., Hamilton, H.B., Otake, M., Goriki, K., Kageoka, T., Fujita, M., Neriishi, S. and Asakawa, J.: Search for mutations affecting protein structure in children of atomic bomb survivors: Preliminary Report. Proc. Natl. Acad. Sci. USA 77: 4221-4225, 1980b.

Neel, J.V., Geroshwitz, H., Mohrenweiser, H.W., Amos, B., Kostyu, D.D., Salzano, F.M., Mestriner, M.A., Lawrence, D., Simoes, A.L., Smouse, P.M., Oliver, W.J., Spielman, R.S. and Neel, J.V. Jr.: Genetic studies on Ticuna, an enigmatic tribe of Central Amazonas. Ann. Hum. Genet. 44: 37-54, 1980c.

Neel, J.V. and Rothman E.D.: Indirect estimates of mutation rates in tribal Amerindians. Proc. Natl. Acad. Sci. USA 75: 5585-5588, 1978.

Neel, J.V. and Thompson, E.A.: Founder effect and number of private polymorphisms observed in Amerindian tribes. Proc. Natl. Acad. Sci. USA 75: 1904-1908, 1978.

Neel, J.V., Ueda, N., Satoh, C., Ferrell, R.E., Tanis, R.J. and Hamilton, H.B.: The frequency in Japanese of genetic variants of 22 proteins V. Summary and comparison with data on Caucasians from British Isles. Ann. Hum. Genet. 41: 429-441, 1978.

Neel, J.V. and Weiss, K.: The genetic structure of a tribal population, the Yanomama Indians XII. Biodemographic studies. Am. J. Phys. Anthrop. 42, 25-52, 1975.

Nei, M.: Extinction time of deleterious mutant genes in large populations. Theor. Pop. Biol. 2: 419-425, 1971.

Nei, M.: *Molecular Population Genetics and Evolution.* Amsterdam and Oxford, North-Holland Publishing Co., 1975.

Nei, M.: Comments on "The intensity of selection for electrophoretic variants in natural populations of *Drosphila*" by B.D.G. Latter, in *Population Genetics and Ecology*, edited by Karlin, S. and Nevo, E., New York, Academic Press, p.409, 1976.

Nei, M.: Estimation of mutation rates from rare protein variants. Am. J. Hum. Genet. 29: 225-232, 1977.

Nei, M. and Chakraborty, R.: Genetic distance and electrophoretic identity of proteins between taxa. J. Mol. Evol. 2: 323-328, 1973.

Nei, M. and Chakraborty, R.: Electrophoretically silent alleles in a finite population. J. Mol. Evol. 8: 381-385, 1976.

Nei, M., Fuerst, P.A. and Chakraborty, R.: Testing the neutral mutation hypothesis by distribution of single locus heterozygosity. Nature 262: 491-493, 1976a.

Nei, M., Chakraborty, R. and Fuerst, P.A.: Infinite allele model with varying mutation rate. Proc. Natl. Acad. Sci. USA 73: 4164-4168, 1976b.

Nei, M., Fuerst, P.A. and Chakraborty, R.: Subunit molecular weight and genetic variability of proteins in natural populations. Proc. Nat. Acad. Sci. USA 75: 3359-3362, 1978.

Nei, M. and Imaizumi, Y.: Estimation of mutation rate in rare recessive traits. Am. J. Hum. Genet. 15: 90-95, 1963.

Nei, M. and Li, W-H.: The transient distribution of allele frequencies under mutation pressure. Genet. Res. 28: 205-214, 1976.

Nei, M. amd Roychoudhury, A.K.: Genetic relationship and evolution of human races. Ann. Hum. Genetics, in Press, 1981.

Nelson, R.L. and Harris, H.: The detection of mutation in human diploid fibroblasts after mutagen treatment using non-selective cloning and enzyme electrophoresis. Mut. Res. 50: 277-283, 1978.

Neyman, J.L.: *The Ecological Basis for Subsistence Change among the Sandawe of Tanzania.* Washington, D.C., National Academy of Sciences, 1970.

Neyman, J.L.: Place and ethnicity among the Sandawe of Tanzania, in *Ethnicity in Modern Africa,* edited by du Toit, B.M., Boulder, Colorado, Westview Press, 1978.

Nurse, G.T.: Ethnic point positions for use in the construction of frequency and distribution maps of Southern Africa. Bull. Int. Comm. Urg. Anthrop. Ethnol. Res. 14: 43-63, 1972.

Nurse, G.T.: The survival of the Khoisan Race. Bull. Int. Comm. Urg. Anthrop. Ethnol. Res. 19: 39-46, 1977.

Nurse, G.T., Bodmer, J.G., Bodmer, W.F., van Leinmen, A., van Rood, J.J., du Toit, E.D. and Botha, M.C.: A reassessment of the HL-A system in Khoisan populations of South West Africa. Tissue Antigens 5: 402-414, 1975.

Nurse, G.T., Botha, M.C. and Jenkins, T.: Serogenetic studies on the San of South West Africa. Hum. Hered. 27: 81-98, 1977.

Nurse G.T., Harpending, H. and Jenkins, T.: Biology and the History of Southern African populations, in *Evolutionary Models and Studies in Human Diversity*, edited by Meier, R.J., Otten, C.M. and Abdel-Hameed, F., The Hague and Paris, Mouton Publishers, 1978.

Nurse, G.T. and Jenkins, T.: The Griqua of Campbell, Cape Province, South Africa. Am. J. Phys. Anthrop. 43: 71-78, 1975.

Nurse, G.T. and Jenkins, T.: *Health and the Hunter-Gatherer*, Basel, S. Karger, 1977a.

Nurse, G.T. and Jenkins, T.: Serogenetic studies on the Kavango peoples of South West Africa. Ann. Hum. Biol. 4: 465-478, 1977b.

Nurse, G.T. and Jenkins, T.: Riemvasmaak before resettlement. S.A. Jour. Sci. 74: 339-341, 1978.

Nurse, G.T., Jenkins, T. and Elphinstone, C.D.: Mseleni Joint disease: Population genetic studies. S.A. Jour. Sci. 70: 360-365, 1974.

Nurse, G.T., Jenkins, T., Santos David, J.H. and Steinberg, A.G.: The Njinga of Angola: a serogenetic study. Ann. Hum. Biol. 6: 337-348, 1979.

Nurse, G.T., Lane, A.B. and Jenkins, T.: Serogenetic studies on the Dama of South West Africa. Ann. Hum. Biol. 3: 33-50, 1976.

Nyhan, W.: Genetically determined molecular variation in man. TIBS 2: 121-122, 1977.

O'Callaghan, M.: *Namibia: the Effects of Apartheid on Culture and Education*, Unesco, 1977.

Ohta, T.: Population size and rates of evolution. J. Mol. Evol. 1: 305-314, 1972.

Ohta, T.: Statistical analyses of *Drosophila* and human protein polymorphisms. Proc. Natl. Acad. Sci. USA 72: 3194-3196, 1975.

Ohta, T.: Role of very slightly deleterious mutations in molecular evolution and polymorphism. Theor. Pop. Biol. 10: 254-275, 1976.

Ohta, T.: In *Molecular Evolution and Polymorphism*, edited by Kimura, M., Mishima, National Institute of Genetics, Japan, pp. 148-167, 1977.

Ohta, T. and Kimura, M.: A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet. Res. 22: 201-204, 1973.

Papiha, S.S., Bernal, J.E., Roberts, D.F., Habeebullah, C.M. and Mishra, S.C.: C'3 polymorphism in some Indian populations. Hum. Hered. 29: 193-196, 1979.

Pingle, U.: A comparative study of mating systems and marriage distance patterns between five tribal groups of Utnur Taluka, Adilabad District of Andhra Pradesh. Proc. 2nd Ann. Conf. Indian Soc. Hum. Genet., Calcutta, pp. 1-15, 1975.

Pious, D. and Soderland, C.: HL-A variants of cultured human lymphoid cells: evidence for mutational origin and estimation of mutation rate. Science 197: 769-771, 1977.

Raa, E.T.: The couth and the uncouth: ethnic, social and linguistic divisions among the Sandawe of central Tanzania. Anthropos 65: 127-153, 1970.

Radcliffe-Brown, A.R.: Former numbers and distribution of the Australian Aborigines. *Official Yearbook of the Commonwealth of Australia*, 23: 671-696, 1930.

Ramesh, A., Blake, N.M., Vijaykumar, M. and Murty, J.S.: Genetic studies on the Chenchu tribe of Andhra Pradesh, India. Hum. Hered. 30: 291-298, 1980.

Ramesh, A., Murty, J. and Blake, N.M.: Genetic studies on the Kolams of Andhra Pradesh, India. Hum. Hered. 29: 147-153, 1979.

Rao, P.M., Blake, N. and Veerraju, P.: Genetic studies on the Savara and Jatapu Tribes of Andhra Pradesh. Hum. Hered. 28: 122-131, 1978.

Rao, P.R. and Goud, J.D.: Sickle-cell haemoglobin and glucose-6-phosphate dehydrogenase deficiency in tribal populations of Andhra Pradesh. Indian J. Med. Res. 70: 807-813, 1979.

Rao, P.R., Goud, J.D. and Swamy, B.R.: Serum albumin variants from populations of Andhra Pradesh, S. India. Hum. Genet. 51: 221-224, 1979.

Roberts, D.F., Papiha, S.S., Rao, G.N., Habeebullah, C.M., Kumar, N. and Murty, K.J.R.: A genetic study of some Andhra Pradesh populations. Ann. Hum. Biol. 7: 199-212, 1980.

Rothman, E.D. and Adams, J.: Estimation of expected number of rare alleles of a locus and calculation of mutation rate. Proc. Natl. Acad. Sci. USA 75: 5094-5098, 1978.

Saha, N., Kirk, R.L., Shanbhag, S., Joshi, S.H. and Bhatia, H.M.: Genetic studies among the Kadar of Kerala. Hum. Hered. 24: 198-218, 1974.

Saha, N., Kirk, R.L., Shanbhag, S., Joshi, S.R. and Bhatia, H.M.: Population genetic studies in Kerala and the Nilgiris (South West India) Hum. Hered. 26: 175-197, 1976.

Santachiara-Benerecetti, A.S., Beretta, M., Negri, M., Ranzani, G., Antonini, G., Barberio, C., Modiano, G. and Cavalli-Sforza, L.L.: Population genetics of red cell enzymes in Pygmies: A conclusive account. Am. J. Hum. Genet. 32: 934-954, 1980.

Santachiara-Benerecetti, A.S., Cattaneo, A. and Meera Khan, P.: A new variant allele AK^5 of the red cell adenylatekinase polymorphism in a non-tribal Indian population. Hum. Hered. 22: 171-173, 1972a.

Santachiara-Benerecetti, S.A., Cattaneo, A. and Meera Khan, P.: Rare phenotypes of PGM_1 and PGM_2 loci and a new PGM_2 variant allele in the Indians. Amer. J. Hum. Genet. 24: 680-685, 1972b.

Santachiara-Benerecetti, A.S., Ranzani, G.N. and Antonini, G.: Studies on African Pygmies V. Red cell acid phosphatase polymorphism in Babinga Pygmies: high frequency of ACP^R allele. Am. J. Hum. Genet. 29: 635-638, 1977.

Schapera, I.: *The Tswana.* London, International African Institute, 1953.

Schull, W.J., Ferrell, R.E. and Rothhammer, F.: Genes, enzymes and hypoxia, in *Ecological Genetics: The Interface*, edited by Brussard, P., New York, Springer, pp. 73-90, 1978.

Schull, W.J. and Neel, J.V.: *The effects of inbreeding in Japanese children.* New York, Harper and Row, 1965.

Serjeantson, S.: *The Population Genetic Structure of the North Fly River Region of Papua New Guinea.* Ph.D. dissertation, Univ. of Hawaii, 1970.

Serjeantson, S.: Marriage patterns and fertility in three New Guinean populations. Hum. Biol. 47: 399-413, 1975.

Shaw, C.R.: Electrophoretic variation in enzymes Science 149: 936-943, 1965.

Siciliano, M.J., Bordelon, M.R. and Kohler, P.O.: Expression of human adenosine deaminase after fusion of adenosine deaminase deficient cells with mouse fibroblasts. Proc. Natl. Acad. Sci. USA 75: 936-940, 1978.

Silberbauer, G.B.: *Bushman Survey Report.* Geberones, Bechuanaland Govt., 1965.

Smith, L.R.: *The Aboriginal population of Australia.* Canberra, Australian National University Press, 1980.

Sommer, S.S. and Cohen, J.E.: The size distribution of proteins, mRNA and nuclear RNA. J. Mol. Evol. 15: 37-57, 1980.

Stamatoyannopoulos, G.: Possibilities for demonstrating point mutations in somatic cells, as illustrated by studies of mutant haemoglobins, in *Genetic Damage in Man caused by Environmental Agents*, Edited by Berg, K., New York, Academic Press, pp. 49-62, 1979.

Stamatoyannopoulos, G., Nute, P.E., Papayannopoulou, T., McGuire, T., Lim, G., Bunn, H.F. and Rucknagel, D.: Development of a somatic mutation screening system using Hb mutants IV. Successful detection of red cells containing the human frameshift mutants Hb Wayne and Hb Cranston using monospecific flourescent antibodies. Am. J. Hum. Genet. 32: 484-496, 1980.

Steinberg, A.G., Jenkins, T., Nurse, G.T. and Harpending, H.C.: Gamma globulin groups of the Khoisan peoples of Southern Africa: evidence for polymorphism for a $Gm^{1,5,13,14,21}$ haplotype among the San. Am. J. Hum. Genet. 27: 528-542, 1975.

Steinkraus, W. and Pence, A.: *Languages of the Goilala Sub-District.* Port Moresby, Department of Information and Extension Services, 1964.

Stevenson, A.C. and Kerr, C.B.: On the distribution of frequencies of mutation to genes determining harmful traits in man. Mut. Res. 4: 339-352, 1967.

Stow, G.: *The Native Races of Southern Africa.* London, Swann Sonneyschein, 1905.

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

Takahata, N.: Composite stepwise mutation model under the neutral mutation hypothesis. J. Mol. Evol. 15: 13-20, 1980.

Taylor, A.J.: *Syntax and Phonology of Motu (Papua): A Transformational Approach.* Ph.D. dissertation, Canberra, The Australian National University, 1970.

Tchen, P., Séger, J., Bois, E., Grenand, F., Friboug-Blanc, A. and Feingold, N.: A genetic study of two French Guiana Amerindian populations II rare electrophoretic variants. Hum. Genet. 45: 317-326, 1978.

Templeton, A.R.: The theory of speciation *via* the founder principle. Genetics 94: 1011-1038, 1980.

Territory of Papua New Guinea Bureau of Statistics,: *Papua New Guinea, Summary of Statistics.* Port Moresby, Government Printer, 1971-72.

Thompson, E.A. and Neel, J.V.: Probability of founder effect in a tribal population. Proc. Natl. Acad. Sci. USA 75: 1442-1445, 1978.

Tills, D., Van Den Branden, J.L., Clements, V.R. and Mourant, A.E.: The world distribution of electrophoretic variants of the red cell enzyme adenylate kinase (ATP: AMP phosphotransforase) EC 2.7.4.3. Hum. Hered. 21: 302-304, 1971.

Tindale, N.B.: Tribal and intertribal marriage among the Australian Aborigines. Hum. Biol. 25: 169-190, 1953.

Tindale, N.B.: *Aboriginal Tribes of Australia.* Los Angeles, University of California Press, 1974.

Traut, H.: On the calculation of human mutation rates from changes in sex-ratio. Ann. Hum. Genet. 33: 45-51, 1969.

Trevor, J.C.: The physical characters of the Sandawe. J. Roy. Anthrop. Inst. 77: 61-78, 1947.

Turner, J.R.G., Johnson, M.S. and Eanes, W.F.: Contrasted modes of evolution in the same genome: Allozymes and adaptive changes in *Heliconius*. Proc. Natl. Acad. Sci. USA 6: 1924-1928, 1979.

Van de Kaa, D.J.: *The Demography of Papua New Guinea's Indigenous Population.* Ph.D. dissertation, Canberra, The Australian National University, 1971-72.

Van Zeeland, A.A. and Simons, J.W.I.M.: The use of correction factors in the determination ofmutant frequencies in populations of human diploid skin fibroblasts. Mut. Res. 34: 149-158, 1976.

Veerraju, P.: Consanguinity in tribal communities of Andhra Pradesh. Medical Genetics in India 2: 157-164, 1978.

Veerraju, P., Sudhakar Babu, M., Jaikishan, G., Naidu, J.M. and Blake, N.M.: Genetic studies on the Koya Dora and Konda Kammara Tribes of Andhra Pradesh, India. Hum. Hered. submitted (1981).

Vergnes, H., Gherardi, M., Jaeger, D. and Benabadji, M.: Genetic variants of glucose-6-phosphate dehydrogenase in a Saharian and Pygmy family. Hum. Hered. 29: 50-56, 1979.

Vergnes, H. and Gourdin, D.: Further data on the distribution of some red cell enzyme variants in African populations. Hum. Hered. 24: 463-471, 1974.

Vergnes, H., Sevin, A., Sewin, J. and Jaeger, G.: Population genetic studies of the Aka pygmies (central Africa). Hum. Genet. 48: 343-355, 1979.

Voelker, R.A., Langley, C.M., Leigh-Brown, A.J., Ohnishi, S., Dickson, B., Montgomery, E. and Smith, S.E.: Enzyme null alleles in natural populations of *Drosophila melanogaster*: frequencies in a North Carolina population. Proc. Natl. Acad. Sci. USA 77: 1091-1095, 1980a.

Voelker, R.A., Schaffer, H.E. and Mukai, T.: Spontaneous allozyme mutations in *Drosophila melanogaster*: rate of occurrence and nature of the mutants. Genetics 94: 961-968, 1980b.

Vogel, F.: Spontaneous mutation in man, in *Chemical Mutagenesis in Mammals and Man*, edited by Vogel, F. and Rohrborn, G., New York, Springer, pp. 16-68, 1970.

Vogel, F.: Spontaneous mutations in man. Adv. Hum. Genet. 5: 223-318, 1975.

Vogel, F. and Motulsky, A.G.: *Human Genetics*. Heidelberg, Berlin and New York, Springer-Verlag, 1979.

Vogel, F. and Rothenberg, R.: Spontaneous mutation in man. Adv. Hum. Genet. 5: 223-318, 1975.

Voorhoove, C.L.: The central and western areas of the trans-New Guinea phylum: central and western trans-New Guinea phylum languages, in *New Guinea Area Languages and Language Study*, Vol. I, edited by Wurm, S.A., Ser.C, No. 38, Pacific Linguistics, Canberra, The Australian National University, pp. 345-459, 1975.

Ward, R.D.: Relationship between enzyme heterozygosity and quaternary structure. Biochem. Genet. 15: 123-135, 1977.

Ward, R.D.: Subunit size of enzymes and genetic heterozygosity in vertebrates. Biochem. Genet. 16: 799-810, 1978.

Watson, J.D. and Crick, F.H.C.: The structure of DNA. Cold Spring Harbor Sym. Quant. Biol. 18: 123-131, 1953.

Watson, J.B.: The significance of a recent ecological change in the central highland of New Guinea. J. Polynesian Soc. 74: 438-450, 1965a.

Watson, J.B.: From hunting to horticulture in the New Guinea highlands. Ethnology 4: 295-309, 1965b.

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

Watterson, G.A.: The sampling theory of selectively neutral alleles. Adv. Appl. Prob. 6: 463-488, 1974.

Watterson, G.A.: Heterosis or Neutrality. Genetics 85: 789-814, 1977 .

Watterson, G.A.: The homozygosity test of neutrality. Genetics 88: 405-417, 1978a.

Watterson, G.A.: An analysis of multi allelic data. Genetics 88: 171-179, 1978b.

Watterson, G.A., . . . and Anderson, R.: Detecting natural selection: The stepwise mutation model with heterosis. Austral. J. Statist. 22: 125-142, 1980.

Welch, S.G., Mills, P.R. and Gaensslen, R.E.: Phenotypic distributions of red cell glutamate-pyruvate transaminase (E.C. 2.6.1.2.) isoenzymes in British and New York populations. Hum. Genet. 27: 59-62, 1975.

Wellington, J.H.: *South West AFrica and Its Human Issues.* Oxford, Oxford Univ. Press, 1967.

Winter, R.M.: Estimation of male to female ratio of mutation rates from carrier detection tests in X-linked disorders. Am. J. Hum. Genet. 32: 582-588, 1980.

Wright, S.: Evolution in Mendelian populations. Genetics 16: 97-159, 1931.

Wright, S.: The distribution of gene frequencies under irreversible mutation. Proc. Natl. Acad. Sci. USA 24: 253-259, 1938.

Wright, S.: Isolation by distance in diverse systems of mating. Genetics 31: 39-59, 1946.

Wurm, S.A.: Language distribution in the New Guina area, in *New Guinea area Languages and Language Study,* Vol. I, edited by Wurm, S.A., Canberra, The Australian National University, Ser. C.,No. 38, Pacific Linguistics, pp. 3-38, 1975a.

Wurm, S.A.: Eastern central trans-New Guinea phylum languages, in *New Guinea Area Languages and Language Study,* Vol. I, edited by Wurm, S.A., Ser. C, No. 38, Pacific Linguistics, Canberra, The Australian National University, pp. 461-526. 1975b.

Yasuda, N.: An average mutation rate in man. Jap. J. Hum. Genet. 18: 279-287, 1973.

Yasuda, N. and Kondo, K.: No sex difference in mutation rates of Duchenne muscular dystrophy. J. Med. Genet. 17: 106-111, 1980.

Z'Graggen, J.A.: *Classificatory and Typological Studies in Languages of the Madang District.* Ser. C., No. 19, Pacific Linguistics, Canberra, The Australian National University, pp. 1-179, 1971.

Z'Graggen, J.A.: The north-eastern areas of the trans-New Guinea phylum: The Madang-Adelbert range sub-phylum, in *New Guinea Area Languages and Language Study,* vol. I, edited by Wurm, S.A., Ser. C., No. 38, Pacific Linguistics, Canberra, The Australian National University, pp. 569-612, 1975b.

Z'Graggen, J.A.: Austronesian languages: Madang Province, in *New Guinea Area Languages and Language Study,* Vol. II, edited by Wurm, S.A., Ser. C., No. 38, Pacific Linguistics, Canberra, The Australian National University, pp. 285-300, 1975a.

Zouros, E.: Hybrid molecules and the superiority of the heterozygote. Nature 262: 227-229, 1976.

Zouros, E.: Mutation rates, population sizes and amounts of electrophoretic variation of enzyme loci in natural populations. Genetics 92: 623-646, 1979.

Publications

1. The frequency of private electrophoretic variants in Australian Aborigines and indirect estimates of mutation rate. Am. J. Hum. Genet. 31: 731-740, 1979. (with Drs N.M. Blake and R.L. Kirk).
2. Factors affecting electromorph mutation rates in man: An analysis of data from Australian Aborigines. Ann. Hum. Biol. 7: 45-54, 1980.
3. Rare allele heterozygosity and relative electromorph mutation rates in man. Ann. Hum. Biol. 8: 263-276, 1981.
4. The frequency of private electrophoretic variants and indirect estimates of mutation rate in Scheduled Tribes from South India. Actan 5: 67-87, 1981.
5. Frequency of private electrophoretic variants and indirect estimates of mutation rate in Papua New Guinea. Am. J. Hum. Genet. 33: 112-122, 1981 (with Drs N.M. Blake, S.W. Serjeantson and R.L. Kirk).
6. Genetic studies on some tribes of the Telangana region, Andhra Pradesh, India. Actan 5: 41-56, 1981. (with Drs N.M. Blake, A. Ramesh, M. Vijaykumar and J.S. Murty).

Am J Hum Genet 31:731 – 740, 1979

The Frequency of Private Electrophoretic Variants in Australian Aborigines and Indirect Estimates of Mutation Rate

K. K. BHATIA,[1] N. M. BLAKE, AND R. L. KIRK

SUMMARY

The number of "private" electrophoretic variants of enzymes controlled by 25 loci has been used to obtain estimates of mutation rate in Australian Aborigines. Three different methods yield values of 6.11×10^{-6}, 2.78×10^{-6}, and 12.86×10^{-6}/locus per generation for the total sample of Aborigines. One tribal population of Waljbiri in central Australia gives values of 2.99×10^{-6} and 2.04×10^{-6} for two of the methods, the third being unapplicable. The mean mutation rate for the total Aboriginal sample of 7.25×10^{-6} is very similar to the value obtained by Neel and his colleagues for Amerindians in South America.

INTRODUCTION

Several studies have used the frequency of private electrophoretic variants of blood proteins detected in samples from local human populations to indirectly estimate the average mutation rate per locus in man [1 – 5]. Neel and his colleagues and Tchen et al. based their calculations on data collected by themselves and collaborators from Amerindian populations in South America. Chakraborty and Roychoudhury relied on results published by workers from three laboratories, including our own, for tribal populations in India.

The formulations used and the basic data and assumptions needed in estimating the mutation rate from the frequency of rare variants have been detailed by Neel and Rothman [3]. They conclude that for electrophoretic variants the mutation rate averages 16×10^{-6}/locus per generation in Amerindian populations, but they point out that the possibility exists for variation in mutation rate on an ethnic or regional basis. Since this possibility requires exploration before statements can be made on the

Received March 26, 1979.

[1] All authors: Department of Human Biology, John Curtin School of Medical Research, Canberra, Australia.

© 1979 by the American Society of Human Genetics. 0002-9297/79/3106-0004$01.00

average mutation rate for man as a whole, we are analyzing our own extensive data for populations in the Southwest Pacific and Australian regions. We report here results for the Aboriginal populations of Australia. Results will be published later for populations in Papua, New Guinea, and for other parts of South and Southeast Asia and the Pacific.

THE STUDY POPULATION

At the time of first European contact, the Aborigines were spread across the Australian continent, having exploited, with few exceptions, all the available ecological situations. Their presence in the continent is dated back to at least 40,000 years, though the occupation of the more arid areas in the center probably took place no more than 10,000 years ago [6]. At the time of European contact the population of Aborigines has been estimated at about 250,000 [7], and the population was divided into several hundred tribal and local groups varying in size from 100 to several thousand persons [8].

During the last 200 years the Aboriginal population of Australia fell dramatically, reaching its lowest reported level in the census of 1921. This population decrease was not uniform; in some areas such as Tasmania, the eclipse was total, while in many others across the southern portion of the continent there are few, if any, persons of full Aboriginal descent remaining. In areas more remote from European settlement the decline in numbers was less, but even here the total may have been reduced to 50% before the increase in population characterizing the present situation commenced. The present analysis is based on samples from this area of minimum disturbance shown in figure 1.

There are no accurate records of the age structure in traditional Aboriginal populations. Available data refer to populations already exposed to varying degrees of European contact. At present, the age structure for persons of full Aboriginal descent shows a heavy-based pyramid with only 41.6% in the 15–44 years age group [9]. In the traditional situation, each population may have varied in demographic parameters influenced by natural disasters such as prolonged drought or cyclones. Such factors may have led to drastic reductions in number followed by subsequent population expansion or by replacement through migration from neighboring groups. Over a longer time period, however, we assume that the population of the continent was in equilibrium, and that the average net increase was zero.

Since the precise boundary of the total Aboriginal population in our surveys is difficult to define, we have provided data also for one specific tribal group, defined by the spoken language Waljbiri, one of the largest linguistic groups in the Northern Territory. The Waljbiri territory (see fig. 1) covers 35,000–40,000 square miles of arid and semi-arid country, and the population density averages one person per 25–27 square miles [10].

Meggitt's detailed study of the Waljbiri revealed that the dialectical Waljbiri tribe is further divided into four subgroups, namely, Yalpari (Lander), Waneiga, Walmalla, and Ngalia. Marriages between the subgroups are frequent, according to Meggitt. Tindale [11], however, found only 1.3% marriages between Ngalia and Walmalla, and no Yalpari-Waneiga marriages were recorded. Birdsell [12] claims that before 1935

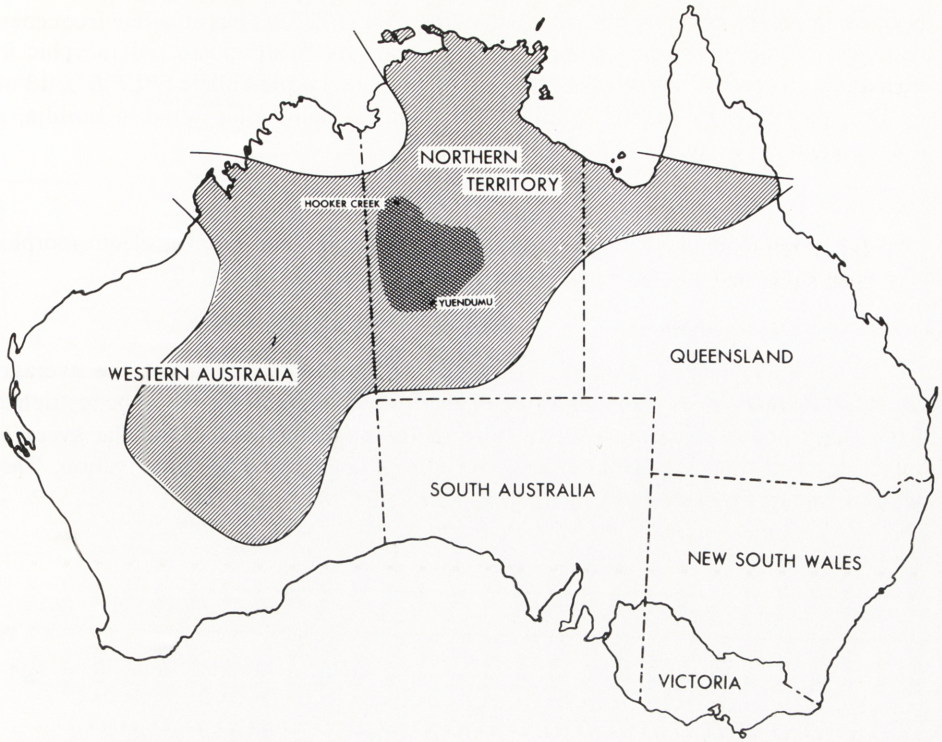

FIG. 1. —Map of Australia showing area sampled (*diagonal hatching*) and tribal territory of the Waljbiri (*cross hatching*).

the Ngalia subgroup was quite distinct from the other Waljbiri. Intertribal marriages involving Ngalia, however, were significantly higher at 6% – 7%.

The Waljbiri in our series were sampled mainly at two localities, Yuendumu and Hooker Creek. The Yuendumu Waljbiri predominantly belong to the Ngalia subgroup, though some reside also at Hooker Creek. Although we have pooled the results for all Waljbiri, our data indicate a clear-cut heterogeneity between the populations at these two localities.

THE LABORATORY DATA

Our analysis is confined to data for red cell enzyme proteins and hemoglobin, representing products of genes at 25 loci. The basic data have been tabulated recently by Blake [13] and are summarized in table 1. A total of 16 detected variants restricted to Australian Aborigines are listed in table 2, together with the number of copies observed and their gene frequencies. Three of the variant alleles ($PGM_2{}^3$, $CA_1{}^9$, and $CA_2{}^4$) have achieved frequencies above 1% and can be classified as polymorphic. Two others (PGD^{Elcho} and $PEP\ B^6$) have allele frequencies approaching 1%, and the remainder are more restricted, the number of copies ranging from one to 14. Table 2 also shows separately the number of rare variants detected in the Waljbiri tribe. Only five of the 16 rare variants among Aborigines were detected among the Waljbiri, four of

these being polymorphic in this tribe, while the other (*PEP* B^6) has an allele frequency of 0.74%. Three of the polymorphic alleles among the Waljbiri are polymorphic in Aborigines in general. In the case of the other peptidase variant allele (*PEP* B^7), 13 of the 14 copies occurred among Waljbiri, the other example being found in Luridja, a group known to intermarry with the Waljbiri.

METHODOLOGY

So far, three methods of indirectly calculating the mutation rate for electromorphs have been suggested and have been reviewed by Neel and Rothman [3].

Kimura and Ohta's Method

In Kimura and Ohta's method [14], three parameters are involved: the average number of mutant alleles per locus (I) estimated from all variants known to be restricted to the study populations; the effective size of the population (N_e), and the average mutant survival time in generations ($\bar{t}_o$) for alleles not moving toward fixation. The mutation rate, μ, is given by

$$\mu = \frac{I}{2N_e} \times \frac{1}{\bar{t}_o} .$$

TABLE 1

GENETIC MARKERS IN AUSTRALIAN ABORIGINES

Locus no.	Enzyme system	Abbreviation	Sample size
1	6-Phosphogluconate dehydrogenase	*6PGD*	4035
2	Acid phosphatase-1	ACP_1	4016
3	Phosphoglucomutase-1	PGM_1	3919
4	Phosphoglucomutase-2	PGM_2	3790
5	Peptidase A	*PEPA*	3034
6	Peptidase B	*PEPB*	3189
7	Carbonic anhydrase-1	CA_1	3751
8	Carbonic anhydrase-2	CA_2	3751
9	Glyoxylase	*GLO*	1290
10	Adenosine deaminase	*ADA*	1437
11	Esterase D	*EsD*	1556
12	Glutamic pyruvic transaminase	*GPT*	1391
13	Hemoglobin-α	$Hb\alpha$	2692
14	Hemoglobin-β	$Hb\beta$	2692
15	Diaphorase	*DIA*	1861
16	Glucose-6-phosphate dehydrogenase	*G6PD*	1014
17	Malate dehydrogenase-2	MDH_2	2964
18	Superoxide dismutase	*SOD*	1795
19	Lactate dehydrogenase-A	LDH_A	4180
20	Lactate dehydrogenase-B	LDH_B	4180
21	Isocitrate dehydrogenase	ICD_s	1226
22	Phosphohexose isomerase	*PHI*	1569
23	Adenylate kinase-1	AK_1	3535
24	Phosphoglycerate kinase	*PGK*	1569
25	Glutamic oxaloacetic acid transaminase	*GOT*	748

NOTE. —Based on Blake, 1979 [13].

TABLE 2

NUMBER AND FREQUENCIES OF PRIVATE VARIANTS IN AUSTRALIAN ABORIGINES

			Total Population		Waljbiri	
	Enzyme	Variant	No. copies	% Gene frequency	No. copies	% Gene frequency
1	6PGD	PGD^{Elcho}	65	0.81	0	0.00
2	PEPA	$PEP\ A^{3}$	1	0.02	0	0.00
3	PEPB	$PEP\ B^{6}$	53	0.83	6	0.74
4	PEPB	$PEP\ B^{7}$	14	0.21	13	1.43
5	PGM_1	$PGM_1{}^{6}$	4	0.05	0	0.00
6	PGM_1	$PGM_1{}^{7}$	1	0.01	0	0.00
7	PGM_2	$PGM_2{}^{3}$	103	1.36	46	5.68
8	PGM_2	$PGM_2{}^{11}$	6	0.08	0	0.00
9	ACP_1	$ACP_1{}^{F}$	1	0.01	0	0.00
10	CA_1	$CA_1{}^{9}$	192	2.53	36	4.44
11	CA_1	$CA_1{}^{10}$	8	0.11	0	0.00
12	CA_2	$CA_2{}^{4}$	166	2.21	14	1.73
13	LDH_B	$LDH_B{}^{SLOW}$	1	0.01	0	0.00
14	LDH_A	$LDH_A{}^{SLOW}$	2	0.02	0	0.00
15	G6PD	$Gd_B{}^{FAST}$	1	0.04	0	0.00
16	PHI	PHI^{4}	1	0.03	0	0.00

Nei's Method

Nei [2] gives a different formulation, where

$$\mu = \frac{I_q}{2N_e} \times \frac{1}{2 \log_e (2nq)} ,$$

and in this case, I_q is estimated from only those variants whose frequency does not exceed 1%, n is the sample size, and q is set at a value of 0.01.

Rothman and Adams' Method

Rothman and Adams [15] give

$$\mu = \frac{\hat{I}}{2N_e} \times [\bar{g}(1) - \Sigma \bar{g}(i) P_{i1}] ,$$

where $\hat{I}$ is the expected number of mutant alleles in the effective population per locus as estimated from the sample, $\bar{g}(1)$ is the estimated proportion of alleles present in only a single copy, $\bar{g}(i)$ is the estimated proportion of variant alleles of i copies, and P_{i1} is the probability for an allele represented by a single copy having i copies in the previous generation. The relationship between I and $\hat{I}$ is given by Rothman and Adams [15].

ESTIMATION OF I, I_q AND $\hat{I}$

For the total Aboriginal population, 25 enzyme loci have been studied. Of these, 13 showed no polymorphism by the standard definition where the least common allele frequency did not exceed 1%, but private variants were detected at four of these loci. Twelve private variants were distributed among eight of the 12 polymorphic loci. The mean sample size (table 3) for the polymorphic loci without private variants (1419) is

TABLE 3

NUMBER OF LOCI WITH AND WITHOUT PRIVATE VARIANTS AND VALUES OF I IN THE TOTAL ABORIGINAL POPULATION

	POLYMORPHIC LOCI			MONOMORPHIC LOCI			TOTAL LOCI		
	With private variants	Without private variants	Total	With private variants	Without private variants	Total	With private variants	Without private variants	Total
No. loci	8	4	12	4	9	13	12	13	25
No. private variants	12	0	12	4	0	4	16	0	16
I/locus	1.50	0.00	1.00	1.00	0.00	0.31	1.33	0.00	0.64
I_a/locus	1.12	0.00	0.75	1.00	0.00	0.31	1.08	0.00	0.52
I/locus	2.07	0.00	1.38	4.93	0.00	1.52	1.80	0.00	1.07
Mean sample size/locus	3686	1419	2930	2736	2120	2309	3369	1904	2607
Mean sample size/variant	2457	...	2930	2736	...	7506	2527	...	4074

significantly lower than those with private variants (3686). In our data, therefore, the probability of detecting private variants among known polymorphic loci increases with sample size.

The data in table 3 clearly show that the value of I varies also with the type of loci (polymorphic or monomorphic). The probability of detecting private variants increases with sample size, which will also influence the value of I.

In Nei's method [2], I_q is calculated only from those private variants which are not polymorphic. The differences for all loci between I and I_q is 0.12 for the total Aboriginal population. For the Waljbiri population, the only private variant ($PEP\,B^7$) is polymorphic. Since it is an exclusive tribal marker for the Waljbiri, we have used it in the calculation of I_q, giving a value of 0.04.

Rothman and Adams' [15] method gives higher values for $\hat{I}$ than for either I or I_q, since the number of private variants is estimated for the total effective population. The difference is most marked for the monomorphic loci, where $\hat{I}$ is almost five times the value of I or I_q.

THE ESTIMATION OF N_e

As explained earlier, with the data available, it is not possible to give a precise estimate of the "effective" population size because of changing reproductive patterns among Aborigines. Here we use the population in the 15–44 age group in the 1961 census year, adjusted for the proportion of the total population in the surveyed area.

The population of full descent Aborigines in Australia in 1961 was 36,137 (18,899 males; 17,238 females), of which 41.6% were in the age cohort 15–44 years [9], and the area surveyed contains approximately 60% of the total full descent population. This gives a value of $N_e = 9{,}160$.

It can be argued that this does not represent the effective population size of Australian Aborigines during most of their stay on the continent. However, indirect evidence suggests the difference in age structure in traditionally oriented societies is not likely to be very different from the value used here. For example, Tindale [8] has recorded approximate age composition for three nomadic bands encountered in the central desert areas. The mean value for the adult composition of these bands is 28.0%. Since this covers the age range 20–40 years, the composition of the 15–44 cohort will not be very different from the 41.6% derived from the 1961 census. In the case of the Waljbiri, we have age estimates for the Yuendumu population [16]. This gives 43.2% for the 15–44 age cohort. From the total Waljbiri population estimate given by Milliken [17], N_e for Waljbiri becomes 1,173.

Another difficulty is that Aboriginal populations have been subject to a series of bottleneck effects due to the operation of various factors. This could cause the loss of a number of private variants which, in turn, will affect the calculation of I, I_q, and $\hat{I}$. The loss of these private variants, however, will be proportional to the decline in population size. On the other hand, private variants which survive the population crash will increase in number during the subsequent population expansion.

ESTIMATION OF $\bar{t}_0$

Kimura and Ohta [14] showed that the mean survival time for a neutral mutation in

generations in terms of effective population size and total population size is given by

$$\bar{t}_o = 2 \frac{N_e}{N} \log_e (2N) ,$$

in a stationary nonsubdivided population with no reproductive death and the progeny size following a Poisson distribution. Kimura and Maruyama [18], however, argue that if the population is subdivided into loose random mating units between which migration occurs, it may be treated approximately as a single random mating unit, disregarding the substructure of the populations.

Applying Kimura and Ohta's formula to the Aboriginal populations, and using the estimates of N_e given above, we obtain values of $\bar{t}_o = 10.7$ and 7.0 generations for the total Aboriginal population, and Waljbiri, respectively. Neel and Rothman [3] however, consider the values of $\bar{t}_o$ calculated by this method as overestimates. The mean survival time can be simulated for each population, and Li and Neel [19] and Li et al. [20] obtained values between 2.3 to 2.8 generations. However, after making concessions for the various factors influencing the population, Li and Neel [19] believe a mean value of 5.7 generations is more appropriate. We shall use this value here.

ESTIMATION OF MUTATION RATES

Mutation rates estimated by each of the three methods listed above, both for the total Aboriginal population surveyed and for the Waljbiri tribal group are given in table 4. The rates vary within a range from 2.78×10^{-6} to 12.86×10^{-6}, with a mean value of 7.25×10^{-6}/locus per generation. In obtaining the values of μ based on Kimura and Ohta's [14] and Nei's [2] methods, we estimated the mean number of variants per locus using the sample size. Neel and Rothman [3], however, base their estimate on the total effective population size. Accordingly, we have also calculated μ with $I = \hat{I}$ and $I_q = \hat{I}$, and the new values become 10.18×10^{-6} and 5.72×10^{-6}/locus per generation, respectively.

Mutation rates estimated from private variants at polymorphic loci are 2.5– 5 times higher than those estimated from the monomorphic loci. This may be a function of the smaller sample sizes for the private variants at monomorphic loci in our sample, which will have reduced the probability of detecting private variants. Eanes and Koehn [21] recently have also drawn attention to the relationship between sample size and detection of rare electrophoretic variants. Neel and Rothman's [3] method, however, yields a higher value of μ when all loci are considered together. The other two methods give values of μ for all loci intermediate between the values for polymorphic and nonpolymorphic loci, while in Neel and Rothman's method, the value of μ for all loci is higher than for either the polymorphic or nonpolymorphic loci.

The values of $\mu = 2.99 \times 10^{-6}$ and 2.04×10^{-6}/locus per generation for the Waljbiri, using Kimura and Ohta's and Nei's methods, are lower than the values obtained for the total Aboriginal sample. These lower values are due to the fact that while five private variants were detected in the Waljbiri, only one is included for calculating I and I_q. The other four are more widely distributed in the Aboriginal population, and it is not possible to assign the original mutants to the Waljbiri.

TABLE 4

MUTATION RATES ($\times 10^6$) IN AUSTRALIAN ABORIGINES ESTIMATED BY VARIOUS METHODS

	Kimura and Ohta's method I	Nei's method I_q	Rothman and Adams' method $\hat{I}$
Polymorphic loci, total Aboriginal sample	9.55	4.01	11.52
Monomorphic loci, total Aboriginal sample	2.96	1.66	2.28
All loci*, total Aboriginal sample	6.11†	2.78†	12.86
Waljbiri sample	2.99	2.04	...

* Mean = 7.25×10^{-6}.

† Following Neel and Rothman [3], the values of μ ($I = I_q = \hat{I}$) were estimated to be 10.18×10^{-6} and 5.72×10^{-6}, respectively.

Neel and Rothman [3] estimated mean mutation rates based on values for 12 Amerindian tribes in South America by each of the same three methods. The unweighted mean for the 12 tribes averaged for the three methods is 16×10^{-6}/locus per generation. The mean value for the Amerindians is more than twice our value for the total Aboriginal sample. However, Neel and Rothman's value of 16×10^{-6} is based on unweighted means for the tribal samples. If it is recalculated using weights based on the effective population sizes, the weighted mean value becomes 7.2×10^{-6}/locus per generation. It is interesting to note that recently Neel and Thompson [22], using a method based on simulation, arrived at a mean mutation rate of 7.0×10^{-6}/locus per generation. These values are very similar to our own based on the total Aboriginal sample. The value for the Waljbiri, of course, is only one-half that for the total Aboriginal sample. Neel and Rothman found a range of values of $0-51 \times 10^{-6}$/locus per generation for their 12 Amerindian tribes. The Waljbiri, therefore, fall within this range and we assume that if data were available for a similar number of tribal populations in Australia, the range of values may also be similar to those for the Amerindians.

Although the indirect estimation of mutation rates using data on private electrophoretic variants has many problems ranging from the technical factors influencing the recognition of rare variants through sampling design to the estimation of I, I_q, $\hat{I}$, and N_e, it is of great interest that data collected in two different laboratories from studies of different populations on two continents have yielded estimates of μ which are so similar. Further studies are in progress in our laboratory which we hope will make possible a further critical evaluation of this approach to the estimation of human mutation rates.

REFERENCES

1. NEEL JV: Private genetic variants and the frequency of mutation among South American Indians. *Proc Natl Acad Sci USA* 70:3311–3315, 1973
2. NEI M: Estimation of mutation rates from rare protein variants. *Am J Hum Genet* 29:225–232, 1977
3. NEEL JV, ROTHMAN ED: Indirect estimates of mutation rates in tribal Amerindians. *Proc Natl Acad Sci USA* 75:5585–5588, 1978

4. CHAKRABORTY R, ROYCHOUDHURY AK: Mutation rates from rare variants of proteins in Indian tribes. *Hum Genet* 43:179–183, 1978
5. TCHEN P, SÉGER J, BOIS E, GRENAND F, FRIBOURG-BLANC A, FEINGOLD N: A genetic study of two French Guiana Amerindian populations II: rare electrophoretic variants. *Hum Genet* 45:317–326, 1978
6. KIRK RL: *Man Adapting: Human Biology of Australian Aborigines*. Oxford, Oxford University Press. In press, 1979
7. RADCLIFFE-BROWN AR: Former numbers and distribution of the Australian Aborigines. *Official Yearbook of the Commonwealth of Australia* 23:671–696, 1930
8. TINDALE NB: *Aboriginal Tribes of Australia*. Los Angeles, University of California Press, 1974
9. Commonwealth of Australia. *Population and Australia*. Canberra, Government Printing Service, 1975
10. MEGGITT MJ: *Desert People*. Sydney, Australia, Angus and Robertson, 1962
11. TINDALE NB: Tribal and intertribal marriage among the Australian Aborigines. *Hum Biol* 25:169–190, 1953
12. BIRDSELL JB: Local group composition among the Australian Aborigines: a critique of the evidence from fieldwork conducted since 1930. *Curr Anthropol* 11:115–142, 1970
13. BLAKE NM: Genetic variation of red cell enzyme systems in Australian Aboriginal populations. *Occasional Papers in Human Biology* 2:39–82, Canberra, Australian Institute of Aboriginal Studies, 1979
14. KIMURA M, OHTA T: The average number of generations until fixation of a mutant allele. *Genetics* 61:763–771, 1969
15. ROTHMAN ED, ADAMS J: Estimation of expected number of rare alleles of a locus and calculation of mutation rates. *Proc Natl Acad Sci USA* 75:5094–5098, 1978
16. MIDDLETON MR, FRANCIS SH: *Yuendumu and its children*. Canberra, Australian Government Publishing Service, 1976
17. MILLIKEN EP: Aboriginal language distribution in Northern Territory, in *Tribes and Boundaries in Australia*, edited by PETERSON N, Canberra, Australian Institute of Aboriginal Studies, 1976
18. KIMURA M, MARUYAMA T: Pattern of neutral polymorphism in a geographically structured population. *Genet Res* 18:125–131, 1971
19. LI FHF, NEEL JV: A simulation of the fate of a mutant gene of neutral selective value in a primitive population, in *Computer Simulation in Human Population Studies*, edited by DYKE B, MACCLUER JW, New York, Academic Press, 1974, 221–240
20. LI FHF, NEEL JV, ROTHMAN ED: A second study of the survival of a neutral mutant in a simulated Amerindian population. *Am Naturalist* 112:83–96, 1978
21. EANES WF, KOEHN RK: Relationship between subunit size and number of rare electrophoretic alleles in human enzymes. *Biochem Genet* 16:971–985, 1978
22. NEEL JV, THOMPSON EA: Founder effect and number of private polymorphisms observed in Amerindian tribes. *Proc Natl Acad Sci USA* 75:1904–1908, 1978

ANNALS OF HUMAN BIOLOGY, 1980, VOL. 7, NO. 1, 45–54

Factors affecting electromorph mutation rates in man: An analysis of data from Australian Aborigines

K. K. BHATIA

Department of Human Biology, John Curtin School of Medical Research, Canberra

Received 20 July 1979

Summary. The constraints of molecular size and structure on the relative magnitudes of electromorph mutation rates as calculated indirectly have been studied using data for Australian Aborigines. The role of sample size in detecting rare electromorphs is important, In addition, subunit size shows a positive and subunit number a negative correlation with mutation rate. The differences in mutation rates were 2–9-fold when calculated for different categories of the data. The importance of physicochemical constraints are discussed.

1. Introduction

During the past decade a number of statistical methods to calculate the mutation rates at cistron level from electrophoretic data, both direct (Mukai 1970, Mukai and Cockerham 1977) and indirect (Kimura and Ohta 1969, Nei 1977, Rothman and Adams 1978), have been developed. The latter methods are based on the detection of private electrophoretic variants in random samples from isolated populations. Some of the assumptions made are: (1) that there is a complete one-to-one correspondence between the incidence and detection of rare electromorphs, (2) that all the rare alleles observed in the population are introduced and maintained through mutation only, and (3) that there is a constancy of mutation rates, on average, over any subset of protein and enzyme loci.

The class of relationship given by (1) is very difficult to evaluate, as estimates of number of alleles depend critically upon sample size (Harris *et al.* 1974, Koehn and Eanes 1978, Eanes and Koehn 1978, Bhatia *et al.* 1979) and upon the resolution of the experimental techniques employed to discriminate allelic variants (Johnson 1977 a). The last point is not trivial, as new techniques suggest that there exists a large reservoir of previously undetected alleles (Johnson 1977 b). Introduction of new alleles by sources other than mutation, e.g., intragenic recombination, was suggested by Watt (1972), Koehn and Eanes (1976) and Strobeck and Morgan (1978).

The assumption included in (3) is the weakest since inter-locus variability in mutation rates has been noted by Nei *et al.* (1976). On the basis of amino acid substitutions in various polypeptide chains, they found this variability to follow the gamma distribution. Zouros (1979) pointed out that over a large range of species only certain types of enzymes occupy the same tail of the distribution, indicating the role of physicochemical features of the molecules and this may explain, in part, the inter-locus variability in mutation rates.

Parameters of genetic variation, like heterozygosity and the number of rare alleles, are affected by a number of factors. For heterozygosity, these include: substrate specificity (Gillespie and Kojima 1968), physiological function (Johnson 1974), quaternary structure (Zouros 1976, Harris *et al.* 1977, Ward 1977) and subunit size (Koehn and Eanes 1977, Nei *et al.* 1978, Brown and Langley 1979). The relationship among these has been demonstrated for both invertebrate and non-human vertebrate

0301–4460/80/0701 0045 $02·00 © 1980 Taylor & Francis Ltd

species. However, Harris *et al.* (1977) have detected lack of correlation between subunit molecular weight and heterozygosity in European human populations. Nei *et al.* (1978) attributed this to the low level of mean heterozygosity in human populations.

In the case of the number of rare alleles, the factors include: effective population size (Ohta 1972, Rothman and Adams 1978) intragenic recombination (Morgan and Strobeck 1979), subunit size (Eanes and Koehn 1978), founder effect (Thompson and Neel 1978), polymorphism (Harris 1975), bottleneck effect (Bhatia *et al.* 1979) and transient distribution of neutral alleles (Nei and Li 1976). The list is by no means exhaustive and a whole set of cause–effect factors, which include the total number of alleles segregating at a locus, mean level of heterozygosity and subunit number, etc., can be included for their role in the introduction and maintenance of rare alleles in a population. Since the estimation of mutation rates by indirect methods depends on the number of rare alleles, it is important to reassess the role of the above factors in determining these rates. In addition, because of the correspondence between molecular weight and mutation rates and the former's role in introducing interlocus variability in mutation rates at the peptide level (Nei *et al.* 1976), it may be relevant also to calculate the mutation rates at base-pair level (Mukai and Cockerham 1977), making cistronic comparisons independent of molecular weight.

2. Materials and methods

Most researchers who have studied the role of variability in mutation rate and heterozygosity restricted themselves to answering only one or two queries, because of the difficulty of controlling all the factors involved. They compensated for the lack of control by increasing the range of species for which results were given. However, an ideal choice for an answer is a subdivided population, distributed over a large geographical area and sampled extensively. The electrophoretic results for Australian Aborigines reviewed by Blake (1979) seem to provide an adequate set of data for analysis. Blake's data as retabulated by Bhatia *et al.* (1979) have been used in the present study. The loci included, arranged into monomers and multimers, and their respective samples sizes are shown in table 1. The multimer loci are all dimers except for the LDH loci. The subunit molecular weights have been taken from the tabulation by Hopkinson *et al.* (1976). A total of 15 multimeric and 10 monomeric loci have been included. In the absence of any direct relationship between subunit number and subunit size (Hopkinson *et al.* 1976), the data for subunit sizes were also pooled together.

The electromorph mutation rates per cistron per generation were calculated by using the methods of Kimura and Ohta (1969) and Nei (1977). The rates at cistron level were then converted to mutation rates per base pair per generation as suggested by Mukai and Cockerham (1977), with only a slight modification. The mutation rates for multimers were computed by subtracting 14% and 28% from the total number of base pairs for dimers and tetramers respectively. This accounts for the amino acid residues involved in surface interactions (Turner *et al.* 1979).

The coefficients of correlation between various parameters were computed by using both the Spearman's rank order non-parametric and Pearson's product moment correlations. Whenever required, the variables were log-transformed to equalize and normalize the variances. The analysis was performed by structuring different classes within each category to equalize or isolate the role of a particular factor.

Enzyme System	Abbreviation	No. of individuals sampled	Subunit size† (daltons)	Total number of alleles	Total number of rare alleles	Heterozygosity ($1-\Sigma x_i^2$)
(*a*) *Multimers*‡						
Haemoglobin-α	Hb-α	2692	15000	1	—	—
Haemoglobin-β	Hb-β	2692	16000	1	—	—
Superoxide dismutase	SOD_A	1795	16000	1	—	—
Glyoxalase	GLO	1290	24000	2	—	0·0380
Esterase D	EsD	1556	28000	2	—	0·1467
Malate dehydrogenase	MDH	2964	35000	1	—	—
Lactate dehydrogenase-A	LDH_A	4180	35000	2	1	0·0002
Lactate dehydrogenase-B	LDH_B	4180	35000	2	1	0·0004
Glutamic oxalacetic acid transaminase	GOT	748	46000	1	—	—
Peptidase A	Pep A	3034	46000	2	1	0·0008
Isocitrate dehydrogenase	IcD_s	1226	48000	1	—	—
Glutamic pyruvic transaminase	GPT	1391	50000	2	—	0·3211
6-Phosphogluconate dehydrogenase	6-PGD	4035	52000	3	1	0·1031
Glucose-6-phosphate dehydrogenase	G-6-PD	1014	53000	2	1	0·0010
Phosphohexose isomerase	PHI	1569	62000	2	1	0·0006
(*b*) *Monomers*						
Acid phosphatase-1	ACP_1	4016	15000	4	1	0·0675
Adenylate kinase-1	AK_1	3535	22000	1	—	—
Carbonic anhydrase-1	CA_1	3751	29000	3	2	0·0516
Carbonic anhydrase-2	CA_2	3751	29000	2	1	0·0425
Diaphorase	DIA	1861	30000	1	—	—
Adenosine deaminase	ADA	1437	34000	2	—	0·0309
Phosphoglycerate kinase	PGK	1569	50000	1	—	—
Phosphoglucomutase-1	PGM_1	3919	51000	4	2	0·2097
Peptidase B	Pep B	3189	55000	3	2	0·0208
Phosphoglucomutase-2	PGM_2	3790	61000	3	2	0·0284

† After Hopkinson *et al.* (1976).
‡ Except LDH loci, which are tetramers, all multimeric loci are dimers.

Table 1. List of proteins and enzymes included in the study and their respective sample sizes, subunit sizes, number of total and rare alleles and expected heterozygosity.

3. Results

Table 2 shows the distribution of loci at which private variants were detected. The data have been classified into two categories, namely multimers and monomers, to avoid the role of functional constraints in influencing other factors. Although the difference is small between multimers and monomers with respect to subunit size (mean values and S.D.s are $37{\cdot}53 \pm 14{\cdot}85$ and $37{\cdot}60 \pm 15{\cdot}48$) and mean expected heterozygosity (0·0408 and 0·0431) respectively, the retention of these divisions is relevant for other comparisons.

Relationship between the number of rare alleles and:

(*a*) *Sample size.* Eanes and Koehn (1978) and Bhatia *et al.* (1979) showed that the efficiency of estimates of mean number of electrophoretic alleles increases with sample size. This was observed also in the present study. The product moment correlation of the total number of alleles as well as the total number of rare alleles with sample size was significantly positive ($r = 0{\cdot}537$, d.f. 23, $P < 0{\cdot}003$ and $r = 0{\cdot}625$, d.f. 23, $P < 0{\cdot}001$ respectively). The relationship showed better correspondence in monomers ($r = 0{\cdot}667$, d.f. 8, $P < 0{\cdot}018$ and $r = 0{\cdot}71$, d.f. 8, $P < 0{\cdot}011$ respectively), but the correlation with multimers was significant for rare alleles only ($r = 0{\cdot}329$, d.f. 13, $P < 0{\cdot}116$ and $r = 0{\cdot}506$, d.f. 13, $P < 0{\cdot}027$).

(*b*) *Total number of alleles.* A significant correlation exists between the total number of alleles and the number of rare alleles because one is included in the other data set. The mean value of Pearson's coefficient for this correlation was significant at the 0·1% level of probability ($r = 0{\cdot}803$, d.f. 23, $P < 0{\cdot}001$). But since it is an analysis of cause–effect relationship, the results can be appreciated better if some variables which affect both of them simultaneously are standardized. The partial correlations by controlling the sample size and mean amount of heterozygosity, individually and combined, yield similar high relationships, although in monomers, controlling by sample size is non-significant.

(*c*) *Heterozygosity.* Since the mean amount of heterozygosity per locus in any population is a function of the total number of alleles, a correlation between the two is to be expected. According to the stepwise mutation model and the intragenic recombination model, the introduction of new alleles will depend upon the frequencies of existing alleles, which is measured by heterozygosity. In the present data, the estimates of mean heterozygosity and its variance are 0·042 and 0·006 respectively. Spearman's rank order correlations for heterozygosity with rare alleles and total number of alleles are significant ($r = 0{\cdot}498$, d.f. 23, $P < 0{\cdot}01$ and $r = 0{\cdot}852$, d.f. 23, $P < 0{\cdot}01$ respectively). The product moment correlation between number of variants and heterozygosity shows a negative correlation (significant at the 1% level of probability) if the values are controlled for total number of alleles. This suggests that the number of rare alleles as a function of heterozygosity or of the total number of alleles, as inferred from the stepwise mutation model, is misleading, particularly for low values of mean heterozygosity.

(*d*) *Subunit number.* Table 2 shows the distribution of loci at which rare alleles were detected in terms of monomeric and multimeric loci. Whereas about 60% of the monomeric loci exhibit the presence of rare alleles, the fraction is 40% in multimers. The number of rare alleles per locus is also much higher in monomers than in multimers (1·00 against 0·40 per locus; table 3). This indicates that a negative association between the number of subunits and rare alleles exists and for the log-transformed variables the present data show a significant negative correlation ($r = -0{\cdot}483$, d.f. 23, $P < 0{\cdot}007$).

Table 2. Means and SDs of sample size, subunit size and heterozygosity at loci with or without rare alleles.

	Type of enzyme	No. of cistrons	Sample size		Subunit size		Heterozygosity	
			Mean	SD	Mean	SD	Mean	SD
Loci with rare alleles	Multimers	6	3002	1403	47167	10720	0·0177	0·0418
	Monomers	6	3736	288	40000	18188	0·0668	0·0709
	Total	12	3369	1039	43583	11028	0·0422	0·0612
Loci without rare alleles	Multimers	9	1817	780	30888	14385	0·0562	0·1104
	Monomers	4	2100	972	34000	11775	0·0077	0·0154
	Total	13	1904	257	31846	13222	0·0413	0·0934
Total	Multimers	15	2291	1188	37533	14875	0·0408	0·0893
	Monomers	10	3082	1036	37600	15479	0·0431	0·0617
	Total	25	2607	1176	37560	14796	0·0417	0·0780

(*e*) *Subunit size.* Eanes and Koehn (1978) obtained significant correlations between the subunit size and total number of alleles at enzyme loci in human populations. Since Harris *et al.* (1977) found no correlation between subunit size and heterozygosity, this suggests a direct relationship between subunit size and the number of rare alleles. In the present data, the correlation between the total number of alleles and subunit size is low, but the number of rare alleles shows a significant relationship ($r=0{\cdot}463$, d.f. 23, $P<0{\cdot}01$). The partial coefficient of correlation between the total number of rare alleles and subunit size is increased significantly when controlled for sample size ($r=0{\cdot}666$, d.f. 22, $P<0{\cdot}001$). It is clear that rare alleles are strongly correlated with subunit size when other factors are standardized.

Effect on mutation rates

From the relationships outlined above, it is obvious that there are several factors which influence the number of rare alleles. I have, therefore, recalculated the electromorph mutation rates from the data for Aborigines following the methods of Kimura and Ohta (1969) and Nei (1977).

Table 3 shows the relationship between the sample size and the estimated average mutation rates. The average number of rare alleles per locus is much higher in sample sizes above 3000 than below 3000 (1·27 against 0·14). This results in a nine-fold difference between these two sample sizes when mutation rates are calculated by the method of Kimura and Ohta (1969), which does not take into consideration the effect of sample size. For the purpose of comparison, three categories of $n \geqslant 3000$, $n < 3000$ and all sample sizes were made. The results show a systematic decrease in mutation rates in these respective categories.

The second important factor which operates to influence the incidence of rare alleles is the presence of polymorphism at a particular locus. The difference between mutation rates at polymorphic and non-polymorphic loci is almost two-fold, indicating the fact that the stepwise mutation model can be invoked to explain these differences (table 4). The difference between the multimer and monomer subgroups could not be given weight because of differences of sample sizes and the incidence of heterozygosity. The results in table 5 show the mutation rates per cistron/generation for three different categories of subunit size, each further sub-divided into multimers and monomers. The

Table 3. Sample size and electromorph mutation rates.

Sample size	Type of enzyme	Mean sample size	Mean subunit size	No. of cistrons	Total no. of rare alleles	μ per cistron ($\times 10^6$)	
						Kimura and Ohta's method	Nei's method
	Multimers	3657	42000	4	4	9·56	6·36
⩾3000	Monomers	3707	37428	7	10	13·68	6·34
	Total	3689	39091	11	14	12·19	6·35
	Multimers	1794	35727	11	2	1·74	1·39
<3000	Monomers	1623	38000	3	—	—	—
	Total	1757	36214	14	2	1·36	1·09
	Multimers	2291	37400	15	6	2·00	3·33
All	Monomers	3082	37600	10	10	11·49	4·64
	All	2607	37480	25	16	5·17	3·60

Table 4. Amount of heterozygosity and electromorph mutation rates in the Australian Aborigines. Note the fluctuations in mean sample sizes.

Type of loci	Proportion of heterozygosity $(1-\Sigma x_i^2)$	Enzyme structure	Mean heterozygosity	Mean sample size	Mean subunit size (daltons)	No. of cistrons	Total no. of rare alleles	μ per cistron ($\times 10^6$)	
								Kimura and Ohta's method	Nei's method
Non-polymorphic	<0·02	Multimers	0·0003	2372	37000	11	5	4·34	3·22
		Monomers	—	2322	34000	3	—	—	—
		All	0·0002	2361	36357	14	5	3·41	2·53
Polymorphic	0·02–0·10	Multimers	0·0380	1290	24000	1	—	—	—
		Monomers	0·0370	3322	37167	6	8	12·74	5·43
		All	0·0371	3032	35286	7	8	10·92	4·75
	0·10–0·30	Multimers	0·1903	2327	43333	3	1	3·19	2·37
		Monomers	0·2097	3919	51000	1	2	20·12	12·52
		All	0·1951	2725	45250	4	3	7·17	5·13
	All	Multimers	0·1522	2068	38500	4	1	2·39	1·84
		Monomers	0·0616	3408	39143	7	10	13·65	6·47
		All	0·0946	2920	38909	11	11	9·56	4·89

Table 5. Subunit size and electromorph mutation rates. Note the fluctuations in mean sample sizes for various categories.

Range of subunit size	Type of enzyme	No. of cistrons	No. of rare alleles	Mean subunit size (daltons)	Mean sample size	μ per cistron ($\times 10^6$)	
						Kimura and Ohta's method	Nei's method
<25000 daltons	Multimers	4	—	17750	2117	—	—
	Monomers	2	1	18500	4016	4·78	3·12
	All	6	1	18000	3776	1·62	1·06
25000–50000	Multimers	8	3	43375	2410	3·63	2·65
	Monomers	5	3	34400	2474	5·74	4·20
	All	13	6	39923	2434	4·40	3·25
>50000 daltons	Multimers	3	3	55667	2206	9·56	7·21
	Monomers	3	6	55667	3633	19·12	6·37
	All	6	9	55667	2919	14·34	6·72

Table 6. Electromorph mutation rates ($\times 10^6$) in Australian Aborigines weighted for sample size, subunit size and proportion of cistron involved in surface interactions.

Weighted by	Kimura and Ohta's method			Nei's method		
	Multimers	Monomers	Total	Multimers	Monomers	Total
Unweighted	3·83	9·56	6·12	3·33	4·64	3·60
Sample size	2·00	11·49	5·17	1·73	3·97	3·28
Sample size + subunit size	1·00	13·30	3·97	0·86	4·61	2·51
Sample size + subunit size + molecular surface interactions	0·83	13·30	3·69	0·71	4·61	2·35

Table 7. Unadjusted electromorph mutation rates per base pair in Australian Aborigines.

Type of enzyme	μ per base pair ($\times 10^8$)	
	Kimura and Ohta's method	Nei's method
Multimers	1·02	0·87
Monomers	4·92	1·70
Total	2·45	1·56

pattern in the three categories is of systematic increase with larger subunit sizes. Multimers have consistently lower mutation rates as compared with monomers, although the mean values of subunit sizes and sample sizes are similar.

4. Discussion

From the observations outlined, it is obvious that the structural constraints and cistron sizes of enzymes, besides the role of sample size, determine to a large extent, the relative magnitudes of electromorph mutation rates. Any comprehensive estimate of

mutation rates for a population will thus have to be weighted for sample size and subunit size. In the present data, weighting by these factors leads to a general reduction in the average mutation rates because of the higher invariant nature of loci with low sample sizes and subunit sizes (table 6). Adjustment for amino acid residues involved in surface interactions in multimers reduces further the average mutation rates. This gives new estimates for μ per cistron per generation in Australian Aborigines as $3{\cdot}69 \times 10^{-6}$ and $2{\cdot}35 \times 10^{-6}$ by the methods of Kimura and Ohta (1969) and Nei (1977) respectively. The differences between monomers and multimers are, however, increased substantially after these modifications.

In principle, the interlocus variability in the mutation rates arising from the various cistron sizes should be minimized if we calculate the mutation rates per base pair per generation rather than per cistron per generation. The estimates of μ per base pair per generation are given in table 7.

Despite the incorporation of modifications necessitated by the physicochemical constraints of the molecules and sample sizes, differences among mutation rates still exist. For example, the relationship between the subunit size and mutation rates does not resolve into a simple linear function. Similarly the differences between the multimeric and monomeric enzymes are increased when adjustments are made for the variation arising from the sample size and subunit size, yet the distinction between polymorphic, monomeric, large-subunit enzymes, and monomorphic, multimeric, small-subunit enzymes is clear cut. This indicates that in making comparisons for electromorph mutation rates among various human populations, the number and type of loci included in the estimations should be taken into account.

Acknowledgement

I would like to thank Dr R. L. Kirk for his valuable suggestions.

References

BHATIA, K., BLAKE, N. M., and KIRK, R. L., 1979, The frequency of private electrophoretic variants in Australian Aborigines and indirect estimates of mutation rates. *American Journal of Human Genetics*, **31**, 731–740.

BLAKE, N. M., 1979, Genetic variation of red cell enzyme systems in Australian Aboriginal populations. *Occasional papers in Human Biology*, **2**, 39–82 (Canberra: Australian Institute of Aboriginal Studies).

BROWN, A. J. L., and LANGLEY, C. H., 1979, Correlation between heterozygosity and molecular weight. *Nature*, **277**, 649–651.

EANES, W. F., and KOEHN, R. K., 1978, Relationship between subunit size and number of rare electrophoretic alleles in human enzymes. *Biochemical Genetics*, **16**, 971–985.

GILLESPIE, J. H., and KOJIMA, K., 1968, The degree of polymorphisms in enzymes involved in energy production compared to that in nonspecific enzymes in two *Drosophila ananassae* populations. *Proceedings of the National Academy of Sciences of the United States of America*, **61**, 582–585.

HARRIS, H., 1975, Multiple allelism and isozyme diversity in human populations. In *Isozymes II. Genetics and Evolution*, edited by E. L. Markert (New York: Academic Press), pp. 131–147.

HARRIS, H., HOPKINSON, D. A., and EDWARDS, Y. H., 1977, Polymorphism and subunit structure of enzymes: A contribution to the neutralist–selectionist controversy. *Proceedings of the National Academy of Sciences of the United States of America*, **74**, 698–701.

HARRIS, H., HOPKINSON, D. A., and ROBSON, E. B., 1974, The incidence of rare alleles determining electrophoretic variants: Data on 43 enzyme loci in man. *Annals of Human Genetics*, **37**, 237–253.

HOPKINSON, D. A., EDWARDS, Y. H., and HARRIS, H., 1976, The distribution of subunit numbers and subunit sizes of enzymes: A study of the products of 100 human gene loci. *Annals of Human Genetics*, **39**, 383–411.

JOHNSON, G. B., 1974, Enzyme polymorphism and metabolism. *Science*, **184**, 28–37.

JOHNSON, G. B., 1977 a, Hidden heterogeneity among electrophoretic alleles. In *Measuring Selection in Natural Populations*, edited by F. B. Christiansen Fenchel (Berlin: Springer-Verlag), pp. 223–244.

JOHNSON, G. B., 1977 b, Isozymes, allozymes and enzyme polymorphisms: Structural constraints on polymorphic variation. In *Isozymes II. Current topics in Biological and Medical Research* (New York, Alan R. Liss, Inc.), pp. 11–19.

KOEHN, R. K., and EANES, W. F., 1976, An analysis of allelic diversity in natural populations of *Drosophila*: The correlation of rare alleles with heterozygosity. In *Population Genetics and Ecology*, edited by A. Karlin and E. Nevo (New York: Academic Press), pp. 377–390.

KOEHN, R. K., and EANES, W. F., 1977, Subunit size and genetic variation of enzymes in natural populations of *Drosophila*. *Theoretical Population Biology*, **11,** 330–341.

KOEHN, R. K., and EANES, W. F., 1978, Molecular structure and protein variation within and among populations. *Evolutionary Biology*, **10,** 39–100.

KIMURA, M., and OHTA, T., 1969, The average number of generations until the fixation of a mutant allele. *Genetics*, **63,** 701–709.

MORGAN, K., and STROBECK, C., 1979, Is intragenic recombination a factor in the maintenance of genetic variation in natural populations? *Nature*, **277,** 383–384.

MUKAI, T., 1970, Spontaneous mutation rates of isozyme genes in *Drosophila melanogaster*. *Drosophila Information Survey*, **45,** 99.

MUKAI, T., and COCKERHAM, C. C., 1977, Spontaneous mutation rate of enzyme loci in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences of the United States of America*, **74,** 2514–2517.

NEI, M., 1977, Estimation of mutation rates from rare protein variants. *American Journal of Human Genetics*, **29,** 225–232.

NEI, M., and LI, W. H., 1976, The transient distribution of allele frequencies under mutation pressure. *Genetical Research*, **28,** 205–214.

NEI, M., CHAKRABORTY, R., and FUERST, P. A., 1976, Infinite allele model with varying mutation rate. *Proceedings of the National Academy of Sciences of the United States of America*, **73,** 4164–4168.

NEI, M., FUERST, P. A., and CHAKRABORTY, R., 1978, Subunit molecular weight and genetic variability of proteins in natural populations. *Proceedings of the National Academy of Sciences of the United States of America*, **75,** 3359–3362.

OHTA, T., 1972, Population size and rates of evolution. *Journal of Molecular Evolution*, **1,** 305–314.

STROBECK, C., and MORGAN, K., 1978, The effect of intragenic recombination of the number of alleles in a finite population. *Genetics*, **88,** 829–844.

ROTHMAN, E. D., and ADAMS, J., 1978, Estimation of expected number of rare alleles of a locus and calculation of mutation rates. *Proceedings of the National Academy of Sciences of the United States of America*, **75,** 5094–5098.

THOMPSON, E. A., and NEEL, J. V., 1978, Probability of founder effect in a tribal population. *Proceedings of the National Academy of Sciences of the United States of America*, **75,** 1442–1445.

TURNER, J. R. G., JOHNSON, M. S., and EANES, W. F., 1979, Contrasted modes of evolution in the same genome: Allozymes and adaptive change in *Heliconius*. *Proceedings of the National Academy of Sciences of the United States of America*, **76,** 1924–1928.

WARD, R. D., 1977, Relationship between enzyme heterozygosity and quaternary structure. *Biochemical Genetics*, **15,** 123–135.

WARD, R. D., 1978, Subunit size of enzymes and genetic heterozygosity in vertebrates. *Biochemical Genetics*, **16,** 799–810.

WATT, W. B., 1972, Intragenic recombination as a source of population genetic variability. *American Naturalist*, **106,** 737–753.

ZOUROS, E., 1976, Hybrid molecules and the superiority of the heterozygote. *Nature*, **262,** 227–229.

ZOUROS, E., 1979, Mutation rates, population sizes and amounts of electrophoretic variation of enzyme loci in natural populations. *Genetics*, **92,** 623–646.

Address correspondence to: K. K. Bhatia, Department of Human Biology, John Curtin School of Medical Research, Box 334, Canberra City, ACT 2601, Australia.

Zusammenfassung. Der Zwang der Molekulargröße und -struktur auf die relative Größe der indirekt berechneten elektromorphen Mutationsrate wurde aufgrund von Daten für australische Eingeborene untersucht. Es wurde gefunden, daß die Rolle der Stichprobengröße bei der Entdeckung seltener Elektromorphen wichtig ist. Außerdem zeigt die Größe der Untereinheiten eine positive und die Zahl der Untereinheiten eine negative Korrelation mit der Mutationsrate. Die Unterschiede der Mutationsrate waren zwei- bis neunfach bei Berechnung für verschiedene Kategorien der Daten. Die Bedeutung physikochemischer Zwänge wird diskutiert.

Résumé. Les contraintes du format et de la structure moléculaires sur la valeur relative des taux de mutation électromorphe calculés indirectement ont été étudiées à partir de données sur les aborigènes australiens. Le rôle de l'effectif de l'échantillon dans la détection d'électromorphes rares est trouvé important. En plus, le format de la sous-unité montre une corrélation positive, et le nombre de sous-unités une corrélation négative avec le taux de mutation. Les différences de taux de mutation étaient de 2 à 9 fois quand ils étaient calculés pour différentes catégories de données. L'importance de contraintes physicochimiques est discutée.

ANNALS OF HUMAN BIOLOGY, 1981, VOL. 8, NO. 3, 263–276

Rare allele heterozygosity and relative electromorph mutation rates in man

K. K. BHATIA

Department of Human Biology, John Curtin School of Medical Research, Canberra, Australia

Received 1 October 1980

Summary. Previous studies of human populations have failed to find a significant relationship between genetic variability, as measured by total heterozygosity, and cistron size, as measured by subunit molecular weight of proteins, but the number of different rare alleles in human populations has been shown to be correlated with subunit size. The present paper examines these relationships further, utilizing data on electrophoretic variants at 27 loci for 12 human populations with a total of 800 000 individual system observations.

The results indicate that, if genetic variability is measured by rare allele heterozygosity instead of total heterozygosity, there is a significant correlation with subunit size. In addition, there are significant differences for rare allele heterozygosity between multimeric and monomeric proteins, the range of variability being less in the multimers (and in the total) than for monomers.

Finally, rare allele heterozygosity has a much bigger range of variability than the range of subunit size. By contrast, the range of rare allele heterozygosity between populations is less than ten-fold, a factor not evident in effective population sizes. Both interlocus and interpopulational estimates of relative electromorph mutation rates (REMR) have been calculated, utilizing the distributions of the number of different rare alleles as well as rare allele heterozygosity. The range of these estimates are much lower than the estimates given by Zouros (1979) using total heterozygosity as input.

1. Introduction

In a previous publication (Bhatia 1980), attention was drawn to the positive correlation between electromorph mutation rate and subunit molecular weight using data for Aboriginal populations in Australia. The same data were used to show a negative correlation with the number of subunits in the functional enzyme and also to illustrate the effect of sample size on the ability to detect electromorphs in the population. The analysis of electromorph mutation rates has now been extended to include data from intensive surveys carried out by several different investigators for a number of major human populations: a total of more than 800 000 single locus tests has been analysed.

Two different strategies have been used in examining the factors influencing electromorph mutation rates. In the first, the relationships of sample heterozygosities, or mean single locus heterozygosities over a set of related populations, are analysed, using both parametric and non-parametric correlation methods. In the second, the analysis is restricted simply to the relationship between the number of different electrophoretic alleles and the size and structure of protein molecules.

Using heterozygosity as a measure of genetic variability, the dependence of neutral mutation rates on subunit molecular weight was demonstrated by Brown and Langley (1979), Turner, Johnson and Eanes (1979), Ward (1978) and Koehn and Eanes (1977, 1978) for various vertebrate and invertebrate populations. This class of relationship, however, was not demonstrated in single species tests of *Colias* (Johnson 1979), *Drosophila* (Johnson 1979, Voelker, Schaffer and Mukai 1980) and man (Harris, Hopkinson and Edwards 1977, Nei, Fuerst and Chakraborty 1978, Bhatia 1980).

0301 4460/81/0803 0263 $02.00 © 1981 Taylor & Francis Ltd

However, in single species tests, Harris *et al.* (1977), Ward (1977) and Bhatia (1980) have shown heterozygosity to be negatively correlated with subunit numbers.

Using the second strategy, a relationship between the average number of different alleles per locus and subunit size was demonstrated by Eanes and Koehn (1978) and Bhatia (1980) in pooled data on human populations and Australian Aborigines respectively. This class of relationship is, however, difficult to evaluate, as the non-parametric estimates of the number of electrophoretic alleles depend critically upon sample size (Nei 1977, Eanes and Koehn 1978, Rothman and Adams 1978, Bhatia 1980) and upon the experimental techniques employed to discriminate allelic variants (Johnson 1977).

Variability in the estimates of mutation rate from protein data, corresponding to the variation in subunit size, has been shown to follow the gamma distribution (Nei, Chakraborty and Fuerst 1976, Fuerst, Chakraborty and Nei 1977, Zouros 1979). Zouros (1979) has used these relationships to generate estimates of relative electromorph mutation rates (REMR) in various natural populations. Using total heterozygosity as input, he found the REMRs to vary more than 500 times over a set of protein loci.

Because of the lack of correlation between heterozygosity and subunit size, extension of Zouros's approach to human data will have only a limited value. Instead the data on rare allele variability may be used to generate relative estimates of mutation rate because of its known dependence on subunit molecular weights. In the present paper, therefore, rare allele variability, expressed both as rare allele heterozygosity as well as the number of rare alleles, is utilized to estimate the REMRs.

2. Materials and methods

In the present study, data on population surveys for electrophoretic variants in 10 major ethnic groups have been included. The surveys on Australian Aborigines, Melanesians, Iranians and South Asian tribal populations are from published and unpublished sources of data in this laboratory. The surveys adopted from other sources are: Amerindians (Neel 1978); Japanese (Neel, Ueda, Satoh, Ferrell, Tanis and Hamilton 1978; GPT data from Ishimoto and Kuwata 1974); English (as compiled by Neel *et al.* 1978; Welch, Mills and Gaensslen 1975 for GPT data); Aymara Indians (Schull, Ferrell and Rothhammer 1978); South African Koisan and Negroid populations (based on work by Professor T. Jenkins and his collaborators and compiled by Bhatia *et al.* in preparation).

We have subdivided data on Melanesians into two linguistic groups, namely Austronesians and non-Austronesians, because of their different origins (Wurm 1975). Rare alleles, assigned on the basis of higher frequency to one language group, have been excluded from the other.

The data have been compiled for 27 protein loci (17 multimers and 10 monomers) and are listed in table 1. The multimeric loci are all dimers except the two LDH loci which are tetramers. Subunit molecular weights are taken from the tabulations of Darnall and Klotz (1975) and Hopkinson, Edwards and Harris (1976).

In the present study, a rare allele has been designated as one with less than 20 copies in 1000 determinations. For each population a separate list of rare alleles was prepared. Rare allele heterozygosity (H_r) is defined here as the number of copies contributed by rare alleles per 1000 determinations. The second parameter, the number of rare alleles (K_r) is simply a count of different rare alleles recovered at each locus. For some purposes K_r is specified per 1000 determinations.

Table 1. Interlocus variability in the frequency of rare alleles and estimates of relative electromorph mutation rates (REMR).

Locus	No. of determinations (A)	Rare alleles: Number (B)	Rare alleles: Copies (C)	Rare allele heterozygosity (H_r) $D = 1000\,C/A$	No. of different rare alleles (K_r) $E = 1000\,B/A$	Relative electromorph mutation rates (REMR): REMR (1) $F = D_i/\sum D_i$	REMR (2) $G = E_i/\sum E_i$
Multimers							
Hb-α	49 191	11	170	3·46	0·224	0·0477	0·0256
Hb-β	49 191	11	39	0·79	0·224	0·0109	0·0256
SOD_1	30 327	2	11	0·36	0·066	0·0050	0·0076
GLO	5 658	0	0	0·00	0·000	0·0000	0·0000
EsD	18 993	3	4	0·21	0·158	0·0029	0·0181
MDH	33 186	8	153	4·61	0·241	0·0635	0·0277
LDH_A	34 886	13	47	1·35	0·373	0·0186	0·0427
LDH_B	34 886	8	71	2·04	0·229	0·0281	0·0262
Hp	38 563	8	9	0·23	0·207	0·0032	0·0237
GOT	8 352	4	4	0·48	0·479	0·0066	0·0548
Pep A	32 853	15	59	1·81	0·457	0·0250	0·0523
ICD_1	21 994	9	14	0·64	0·409	0·0088	0·0468
Pep D	7 669	4	49	6·39	0·522	0·0880	0·0597
GPT	14 769	7	23	1·56	0·474	0·0215	0·0542
6PGD	46 884	18	188	4·01	0·384	0·0552	0·0439
Cp	23 244	14	115	4·95	0·602	0·0682	0·0689
PHI	28 060	25	103	3·67	0·891	0·0505	0·1020
Monomers							
ACP_1	46 855	7	45	0·96	0·149	0·0133	0·0171
AK_1	38 385	2	11	0·29	0·052	0·0040	0·0060
CA_1	26 889	5	15	0·56	0·186	0·0077	0·0213
CA_2	17 502	2	40	2·29	0·114	0·0316	0·0130
PGK	17 700	2	112	6·33	0·113	0·0872	0·0129
PGM_1	49 605	27	96	1·94	0·544	0·0267	0·0623
Pep B	33 310	15	118	3·54	0·450	0·0488	0·0515
PGM_2	48 550	14	304	6·26	0·288	0·0862	0·0330
Alb	30 264	9	276	9·12	0·297	0·1256	0·0340
Tf	38 091	23	192	5·04	0·604	0·0694	0·0691

The relationships between rare allele variability and molecular structure were tested using linear regression methods. Whenever necessary, the variables were log transformed to equalize and normalize the variances. Both Pearson's product moment and Spearman's rank order correlations were computed to test the correspondence between different variables.

3. Results

Table 1 shows the distribution of the total number of rare alleles (B) and total number of copies and sample sizes (n), for the 27 protein loci. Columns D and E of the table show the observed estimates of rare allele heterozygosity (H_r) and number of rare alleles (k_r) per 1000 determinations respectively. The weighted mean subunit sizes, sample sizes and rare allele heterozygosities for various classes of subunit size are shown in table 2.

Interlocus variability

Number of rare alleles (K_r). A total of 266 different rare alleles, with an average recovery of one rare allele for every 3016 determinations, was detected. The range is

Table 2. Subunit size, quaternary structure and rare allele variation in 12 human populations.

Range of subunit size	Type of protein	No. of cistrons	Rare alleles		Mean subunit size	Mean sample size	Rare allele heterozygosity (per 1000 determinations)	No. of rare alleles (per locus)	No. of rare alleles (per 1000 determinations)
			No.	copies					
< 25 000 daltons	Multimers	4	24	220	17 750	33 592	1·637	6·00	0·179
	Monomers	2	9	56	18 500	42 620	0·657	4·50	0·106
	Total	6	33	276	18 000	36 601	1·256	5·50	0·150
25 000 –50 000 daltons	Multimers	10	79	433	41 300	24 609	1·760	7·90	0·321
	Monomers	3	9	167	36 000	20 697	2·690	3·00	0·145
	Total	13	88	600	40 076	23 706	1·946	6·77	0·286
> 50 000 daltons	Multimers	3	57	406	55 667	32 279	4·135	19·00	0·580
	Monomers	5	88	986	65 200	39 964	4·934	17·60	0·440
	Total	8	145	1312	61 625	37 251	4·671	18·12	0·487
Total	Multimers	17	160	1059	38 294	28 156	2·212	9·41	0·334
	Monomers	10	106	1209	47 100	34 715	3·482	10·60	0·305
	Total	27	266	2268	41 555	30 585	2·746	9·85	0·322

from none in 5658 determinations for glyoxalase (GLO) to one in 1122 determinations for phosphohexose isomerase (PHI). Despite a significant correlation between the recovery of rare alleles and sample size ($r = 0{\cdot}544$; $p < 0{\cdot}002$), sampling error is unlikely to explain the failure to recover variants for glyoxalase (GLO). The possibility of testing 5658 individuals without detecting a variant is very low ($P < 0{\cdot}001$).

The mean unweighted number of rare alleles ($\bar{K}_r$) per 1000 determinations in monomers and multimers are $0{\cdot}279 \pm 0{\cdot}088$ and $0{\cdot}349 \pm 0{\cdot}053$ respectively. The difference is statistically insignificant, thereby discounting the role of quaternary structure in introducing new alleles.

There is a significant positive correlation between the number of different rare alleles (K_r) and subunit size (m). The values of r_{Km} for total, multimeric and monomeric loci are shown in table 3. Only 34% of the variability in the number of different rare

Table 3. Correlation coefficients (r) between molecular weight, sample size and parameters of rare alleles and the proportions of variance explained by molecular weight variation (r^2).

Parameter	Type of protein	Sample size		Subunit size	
		r	r^2	r	r^2
Number of rare alleles (K_r)	Multimers	0·5149*	0·2651	0·5118*†	0·2619
	Monomers	0·6269*	0·3930	0·6402*†	0·4099
	Total	0·5441**	0·2960	0·5834***†	0·3403
Rare allele heterozygosity (H_r)	Multimers	0·0678	0·0046	0·4314*	0·1861
	Monomers	−0·1652	0·0273	0·7511**	0·5641
	Total	0·0462	0·0021	0·6411**	0·4110

* $0{\cdot}01 < P < 0{\cdot}05$; ** $0{\cdot}001 < P < 0{\cdot}01$; *** $P < 0{\cdot}001$; † Partial correlations after controlling for sample size are 0·8450***, 0·7590*** and 0·7434*** for multimers, monomers and total proteins respectively.

alleles is explained by variability in subunit size. This proportion rises to 55% when the partial correlations are computed, after controlling for sample size. Considered separately, both multimers and monomers show better correspondence with their respective molecular weights (see table 3). The values of r^2 for multimers and monomers are increased to 74% and 57% respectively, when adjustments are made for sample size as control variable. The estimated parameters for the regression line $y = a + bX$ are: $\hat{a} = 0{\cdot}2402$ and $\hat{b} = 0{\cdot}00023$. The small value of $\hat{b}$ is due to the units used for expressing molecular weights. The scattergram for the values at each locus is shown in figure 1.

Rare allele heterozygosity (H_r). The estimates of rare allele heterozygosity (H_r) are not related to fluctuations in sample size ($r = 0{\cdot}046$; $P > 0{\cdot}410$) or to the number of different rare alleles ($r = 0{\cdot}272$; $P > 0{\cdot}085$) (the rank order correlations for the latter are, however, significant). However, a significant positive correlation does exist with subunit size ($r = 0{\cdot}641$; $P < 0{\cdot}001$) with r^2 explaining more than 41% variability contributed by molecular weight. These results are specially significant in view of the lack of correlation between total heterozygosity and molecular weight in human populations. Both multimers and monomers similarly exhibit significant correlations, although the value of r^2 in multimers is only 18% against 56% for monomers (see table 3).

The unweighted mean values of rare allele heterozygosity ($\bar{H}_r$) are $2{\cdot}70 \pm 0{\cdot}47$, $2{\cdot}15 \pm 0{\cdot}48$ and $3{\cdot}63 \pm 0{\cdot}93$ for total, multimeric and monomeric loci respectively. In contrast with the results derived from the number of rare alleles (K_r) per 1000

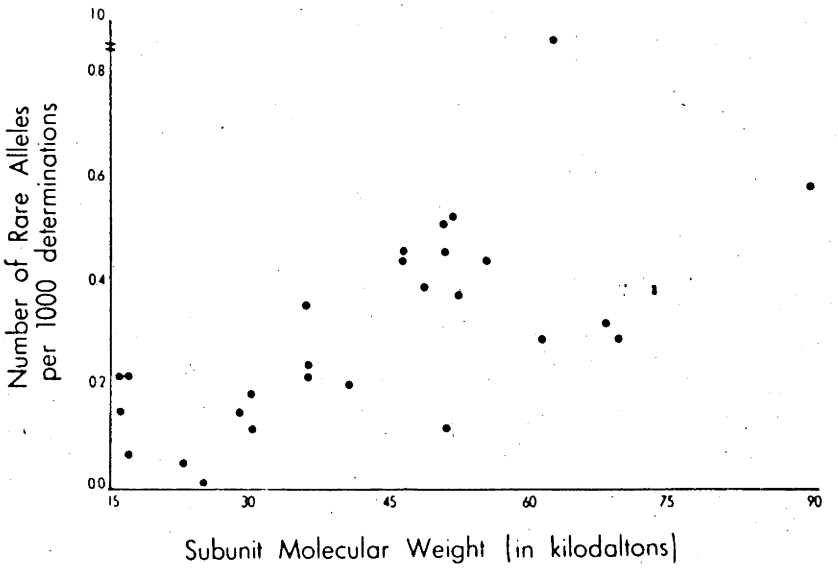

Figure 1. Relationship between the number of different rare alleles per 1000 determinations at a locus and the respective subunit molecular weights in human populations.

determinations, the results for rare allele heterozygosity exhibit significant differences between monomers and multimers. The role of molecular constraints present in multimers in reducing genetic variability is discernible in this parameter.

The scattergram for the values of rare allele heterozygosity (H_r) and subunit molecular weight at each locus is shown in figure 2. The linear regression is

$$\hat{Y} = -0{\cdot}84998 + 0{\cdot}00009\,X$$

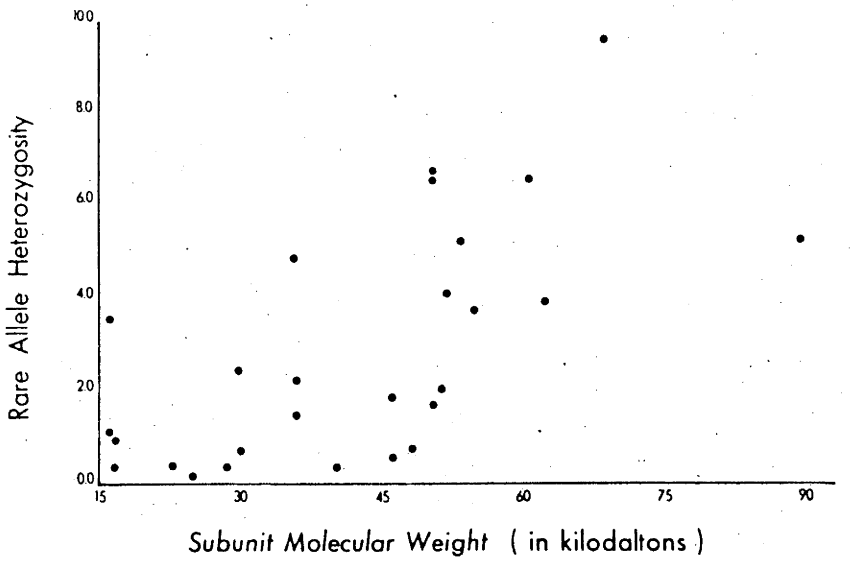

Figure 2. Relationship between the rare allele heterozygosity (copies of rare alleles per 1000 determinations) at a locus and the respective subunit molecular weights in human populations.

Relative electromorph mutation rates (REMR). Two different estimates of relative electromorph mutation rates (REMR) were obtained.

REMR (1) represents the scaled value of rare allele heterozygosity, so that for any locus

$$\text{REMR (1)} = \frac{H_r}{\sum H_r}$$

and REMR (2) represents the scaled value for the number of different rare alleles, so that

$$\text{REMR (2)} = \frac{K_r}{\sum K_r}$$

The values of REMR (1) and REMR (2) are given in the last two columns of table 1. Although both the estimates of REMR show significant postive correlations with subunit molecular weight and have similar rankings, the variability exhibited by the two methods differs widely. After excluding the invariant locus (GLO), the ratios between the lower and upper values for REMR (1) and REMR (2) are 30·1 and 16·7 respectively. The most variant loci are albumin (Alb) for REMR (1) and phosphohexose isomerase (PHI) for REMR (2).

The ratios of REMR (1) and REMR (2) for multimers are 16·00 and 13·42 and for monomers, 30·14 and 10·38 respectively. In comparison, the ratios of minimum to maximum subunit size are only 4·13 and 6·00 in multimers and monomers respectively. Thus the variability in REMRs is 3–5 times more than the observed variability in subunit size.

Interpopulational variability

Since the data on various loci were compiled from different population groups, interpopulation comparison have been made also. Table 4 shows the number of different rare alleles and rare allele heterozygosities in 12 human populations and the corresponding values of REMR (1) and REMR (2).

Table 4. Parameters of rare allele variation and relative electromorph mutation rates in 12 human populations.

Population	No of determinations (A)	Rare alleles: Number (B)	Rare alleles: Copies (C)	Rare allele heterozygosity (H_r) $D = 1000C/A$	No. of different rare alleles (K_r) $E = 1000B/A$	Relative electromorph mutation rates (REMR): (REMR) (1) $F = D_i/\sum D_i$	(REMR) (2) $G = E_i/\sum E_i$
Amerindians	150521	28	426	2·83	0·186	0·0907	0·0393
Japanese	70388	37	175	2·49	0·526	0·0798	0·1112
English	109009	42	83	0·76	0·385	0·0244	0·0814
Australian Aborigines	66258	14	170	2·57	0·211	0·0824	0·0446
Melanesians (Total)	263890	37	992	3·76	0·140	0·1205	0·0296
Austronesians (AN)	89806	17	163	1·82	0·189	0·0584	0·0399
Non-Austronesians (NAN)	174084	19	556	3·19	0·109	0·1023	0·0230
South Asian (Sch. Tribes)	65974	20	178	1·66	0·303	0·0532	0·0640
South African Negroes	39865	15	46	1·15	0·376	0·0369	0·0078
South African Khoisan	16895	7	84	4·98	0·414	0·1596	0·0875
Aymaras	32004	14	73	2·28	0·437	0·0731	0·0924
Iranians	10993	16	41	3·72	1·455	0·1192	0·3075
Total sample (except AN, NAN)	825797	266	2628	2·75	0·322	—	—

Number of rare alleles (K_r). There is a large amount of variability in the detection of rare alleles among different populations. For example, in Iranians a new rare variant was detected for every 687 determinations, whereas in non-Austronesians a rare allele was recovered for every 9162 determinations. Although the detection of rare variants is a logarithmic function of sample size, it is interesting to note that there exists a negative correlation between the sample size and number of different rare alleles per 1000 determinations ($r = -0{\cdot}560$; $P < 0{\cdot}029$). At present it is difficult to explain this result.

As shown above, the recovery of rare alleles for the total population does not differ significantly between monomers and multimers. However, although the values are significant for the individual populations of Japanese, English, Australian Aborigines, South Asian tribes, South African Negroes and Aymaras, in the English and South African Negroes, multimers show more rare variants; momomers are in excess in the other four (table 5).

Rare allele heterozygosity (H_r). Significant heterogeneity in the interpopulation variability of rare allele heterozygosity (H_r) was detected over individual loci except GLO (no variant recovered), EsD and GOT. Coincidentally, the recovery of rare alleles at these loci is, in general, rather low. Chi-square heterogeneity over populations is also found to be significant for weighted mean values for multimeric, monomeric and all loci combined. The mean values of rare allele heterozygosity (H_r) in different populations range from 0·76 in English to 4·98 in South African Khoisans.

It has been pointed out previously that there is a significant difference in rare allele heterozygosity between monomers and multimers for the total sample. The same effect is apparent when considering individual populations, significant differences being present between monomers and multimers for 7 out of 12 populations. In the two African populations, however, the rare allele heterozygosity is significantly higher in multimers. The reason for these differences is not clear.

Relative electromorph mutation rates (*REMRs*). Table 4 shows the range of interpopulational estimates of REMR (1) and REMR (2). The range of REMR(1) is less than an order of magnitude; REMR (2) shows slightly more variation. The differences are smaller in comparison with the effective population sizes of the groups in question.

Table 5. A comparison of rare allele heterozygosity (H_r) and number of different rare alleles (K_r) between monomers and multimers for 12 human populations and total samples.

Population	Rare allele heterozygosity			Number of rare alleles per 1000 determinations	
	Multimers	Monomers	χ^2	Multimers	Monomers
Amerindians	2·53	3·19	5·76*	0·192	0·178
Japanese	1·30	4·20	57·57**	0·433	0·659
English	0·78	0·74	0·04	0·453	0·316
Australian Aborigines	1·99	3·27	10·54**	0·166	0·267
Melanesians (total)	2·41	6·00	214·00**	0·152	0·121
Austronesians	1·95	1·58	1·52	0·193	0·187
Non-Austronesians	1·99	5·14	128·33**	0·130	0·076
South Asian (Sch. Tribes)	2·49	1·15	1·80	0·244	0·401
South African Negroes	1·83	0·46	16·26**	0·593	0·153
South African Khoisan	7·46	1·96	25·61**	0·432	0·392
Aymaras	2·18	2·46	0·26	0·396	0·509
Iranians	4·36	2·68	1·95	1·598	1·217
Total sample	2·21	3·48	143·58**	0·249	0·238

* $0{\cdot}01 < P < 0{\cdot}05$; ** $P < 0{\cdot}01$.

No correspondence between monomers and multimers was detected for REMR (1) across 12 populations ($r = -0{\cdot}036$; $P > 0{\cdot}485$). The results indicate lack of any systematic pressure on the frequency of mutation rates in different human populations, with regard to protein structure.

4. Discussion

From the results outlined above, it is apparent that in human populations, sampled on an adequate scale, the size of molecules, and whether the intact molecule consists of a single subunit or of subunits combined together into multimers, has an important influence on the relative magnitudes of electromorph mutation rates. The range of mean rare allele heterozygosity for these different categories is 3–4 times for both multimers and for all loci, when molecules with similar subunit molecular weights are compared. Monomers, on the other hand, reveal a larger variability, although the range is still less than ten-fold (see table 2).

Analysis of variance among categories reveals that this variability is real rather than stochastic ($F = 6{\cdot}96$; $P < 0{\cdot}01$) but the various parameters of rare allele variation, when normalized for subunit size, indicate non-significant variation among different categories. The molecules, in the middle range of subunit size, however, reveal least variability.

Some of the results in the present study are at variance with our previous analysis of data on Australian Aborigines (Bhatia 1980). Differences in mutation rate of more than an order of magnitude between multimers and monomers, after weighting for sample size, subunit size and adjusting for the proportion of amino acids involved in molecular surface interactions in multimers, in the data on Australian Aborigines are not seen in the present data. The simplicity of proportionality between the molecular size and heterozygosity assumed in this and the previous paper is, however, questionable, especially when individual amino acids, nucleotides and sites within cistrons are known to show variability in their substitution rates (Dayhoff, Schwartz and Orcutt 1978, Kimura 1979, Go and Miyazawa 1980).

Although the magnitude of variability of the relative electromorph mutation rates estimated from rare allele heterozygosity and number of different rare alleles is much smaller than that found by Zouros (1979) using total heterozygosity, the differences between the ranges of subunit size and REMRs are still significant. It appears that, since rare alleles are less likely to be operated upon by systematic negative or positive selection, comparatively large numbers of rare alleles may be maintained by slightly deleterious mutations (Ohta 1976, Li 1978, 1979 a) or bottle-neck effect (Nei 1976, Nei and Li 1976). Although Bhatia (1980) and Chakraborty, Fuerst and Nei (1980) found no correlation between the number of different rare alleles and total heterozygosity, the effect of intragenic recombination (Strobeck and Morgan 1978, Morgan and Strobeck 1979) on loci with unusually high mutation rates may be another factor contributing to this variability. The larger variability seen in the size of mRNAs (Sommer and Cohen 1980) may also decide eventually the total amount of mutations obtained at a particular locus.

The range of interlocus variability in the estimates of H_r is, in general, much higher within populations than in the aggregate data. For example, Harris (1978) has recorded a 150-fold range in the values of H_r in English populations. The data presented here shows similar ranges in other major world populations. Genetic drift and geometric distributions of the copies of rare alleles (Rothman and Adams 1978) are two of the reasons which can be invoked to explain this much larger variability. It may be relevant

to point out here that only 6 out of 12 populations show significant correlations between the molecular weight and rare allele heterozygosity, which indicates clearly that the relationship cannot be demonstrated unequivocally at the level of individual populations. Besides, recent fluctuations in population sizes may also affect these individual population correlations (Li 1979 b).

The distribution of REMRs does not give a good fit to either a gamma or log normal distribution. Cavalli-Sforza and Bodmer (1971) and Yasuda (1973) have examined mutation rates using a log normal and a gamma distribution respectively. Nei *et al.* (1976), Fuerst *et al.* (1977), Chakraborty, Fuerst and Nei (1978), Zouros (1979) and Chakraborty *et al.* (1980) have shown a gamma distribution of mutation rates in proteins supported with similar evidence from distribution of protein subunit sizes. Sommer and Cohen (1980), however, found that the frequency distribution of subunit molecular weights is well described by a log normal distribution. While the subunit molecular weights of the loci included in the present study do, as shown by non-significant values of Pearson's statistics for their log values, follow log normal distribution, the results for REMRs are not so well described by this distribution. One possibility is that compound distributions, which may arise from substitution processes at nucleotide level and distribution of cistron sizes, are involved.

It is interesting to consider if there exists any interpopulational correspondence in the single locus estimates of rare allele heterozygosity. Since the amount of normalized identity (I) between any two distinct populations is negligible for rare alleles, the existence of such correlations must be a function of slightly deleterious mutations (Ohta 1976) or variable mutation rates (Chakraborty *et al.* 1978). The significance of this correlation can be tested using normal tests, since the value of r follows a normal distribution for $I = 0$ (Chakraborty *et al.* 1978). The existence of such a relationship can be seen in the significant correlations between single-locus rare allele heterozygosities of two samples obtained from the same Japanese population (Neel, Satoh, Hamilton, Otake, Goriki, Kageoka, Fujita, Neriishi and Asakawa 1980). In the present study 13 of the 66 possible estimates of the coefficient of correlation for single-locus rare allele heterozygosities among 12 populations are significant. Since half of the populations show significant correlations between rare allele heterozygosity and molecular weight, from the foregoing discussion it could be expected that 15 of the 66 pairwise comparisons will show significant correlations. The close approximation of the observed and expected number of significant correlations is encouraging.

Eanes and Koehn (1977), Chakraborty and Fuerst (1979) and Chakraborty *et al.* (1980) suggest that the correlation between the number of different alleles and molecular weights is generally higher than the correlation between molecular weight and heterozygosity. Chakraborty *et al.* (1980) confirm theoretically that this is expected to be so. For large sample sizes, they expect these correlations to be higher because of the inclusion of slightly deleterious mutations. For rare alleles, over sufficiently large sample sizes, this study shows the results to be otherwise for monomers and the total number of loci (table 3). The partial correlations, after controlling the sample size, however, confirm the observations of Eanes and Koehn (1977) and Chakraborty *et al.* (1980).

The magnitude of interpopulation variability in REMRs recorded in the present study is smaller than the interlocus variability. One of the factors influencing this is the amount of variability compressed within electromorphs which is related to $N\mu$ (Chakraborty and Nei 1976, Nei and Chakraborty 1976). Zouros (1979) has considered these differences to be the relative estimates of effective population sizes (N_e). However,

demographic features of human populations have altered so much in the past and the errors involved in computing estimates of N_e are so large that I prefer to call these estimates interpopulational REMRs rather than relative effective population sizes (REPS).

The need for both absolute direct and indirect estimates of mutation rates in man from proteins has been emphasized by a number of workers (Neel 1973, 1977, Neel and Rothman 1978, Nei 1977, Chakraborty and Roychoudhury 1978, Dudinin and Altukhov 1979, Tchen, Séger, Bois, Grenand, Fribourg-Blanc and Fiengold 1978, Bhatia, Blake and Kirk 1979, Bhatia, Blake, Serjeantson and Kirk 1981, Bhatia 1980). But it will be quite some time before reliable estimates are generated. With the accumulating evidence for the correlation of subunit size and molecular structure with mutation rates in animal and plant species, and now in man, estimates of relative electromorph mutation rates can be extrapolated to real problems in population genetic theory.

Acknowledgements

I am grateful to Dr R. L. Kirk for the help received in the preparation of the manuscript and Dr N. M. Blake with the compilation of data. My thanks are due to Mrs Robbie Williams for typing the manuscript.

References

Bhatia, K., 1980. Factors affecting electromorph mutation rates in man: An analysis of data from Australian Aborigines. *Annals of Human Biology*, **7**, 45–54.

Bhatia, K., Blake, N. M., and Kirk, R. L., 1979. The frequency of private electrophoretic variants in Australian Aborigines and indirect estimates of mutation rate. *American Journal of Human Genetics*, **31**, 731–740.

Bhatia, K., Blake, N. M., Serjeantson, S. W., and Kirk, R. L., 1981. The frequency of private electrophoretic variants and indirect estimates of mutation rate in Papua New Guinea. *American Journal of Human Genetics*, **33**, 112–122.

Brown, A. J. L., and Langley, C. H., 1979. Correlations between heterozygosity and molecular weight. *Nature*, **277**, 649–651.

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

Chakraborty, R., and Fuerst, P. A., 1979. Some sampling properties of selectively neutral alleles. *Genetical Research (Cambridge)*, **34**, 253–267.

Chakraborty, R., Fuerst, P. A., and Nei, M., 1978. Statistical studies on protein polymorphism in natural populations II. Gene differentation between populations. *Genetics*, **88**, 367–390.

Chakraborty, R., Fuerst, P. A., and Nei, M., 1980. Statistical studies on protein polymorphism in natural populations III. Distribution of allele frequencies within populations. *Genetics*, **94**, 1039–1063.

Chakraborty, R., and Nei, M., 1976. Hidden genetic variability within electromorphs in finite populations. *Genetics*, **84**, 385–393.

Chakraborty, R., and Roychoudhury, A. K., 1978. Mutation rates from rare variants of proteins in Indian tribes. *Human Genetics*, **43**, 179–183.

Darnall, D. W., and Klotz, I. M., 1975. Subunit constitution of proteins: A table. *Archives of Biochemistry and Biophysics*, **166**, 651–682.

Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C., 1978. A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*, edited by M. O. Dayhoff (Washington: National Biomedical Research Foundation) Vol. 5, Supp. 3, pp. 345–352.

Dubinin, N. P., and Altukhov, Yu. P., 1979. Gene mutations (*de novo*) found in electrophoretic studies of blood proteins of infants with anomalous development. *Proceedings of the National Academy of Sciences of the United States of America*, **76**, 5226–5229.

Eanes, W. F., and Koehn, R. K., 1977. The correlation of rare alleles with heterozygosity: determination of the correlation for neutral models. *Genetical Research (Cambridge)*, **29**, 223–230.

Eanes, W. F., and Koehn, R. K., 1978. Relationship between subunit size and number of rare electrophoretic alleles in human enzymes. *Biochemical Genetics*, **16**, 971–985.

Fuerst, P. A., Chakraborty, R., and Nei, M., 1977. Statistical studies on protein polymorphism in natural populations. I. Distribution of single locus heterozygosity. *Genetics*, **86**, 455–483.

GO, M., and MIYAZAWA, S., 1980, Relationship between mutability, polarity and exteriority of amino acid residues in protein evolution. *International Journal of Peptide Research*, **15**, 211–224.

HARRIS, H., 1978, Mutation and enzyme variation in human populations. In *Mutations, Biology and Society*, edited by D. N. Walcher, N. Ketchner and H. C. Barnett (New York: Mason Publishing House) pp. 77–98.

HARRIS, H., HOPKINSON, D. A., and EDWARDS, Y. H., 1977, Polymorphism and the subunit structure of enzymes: A contribution to the neutralist-selectionist controversy. *Proceedings of the National Academy of Sciences of the United States of America*, **74**, 698–701.

HOPKINSON, D. A., EDWARDS, Y. H., and HARRIS, H., 1976, The distribution of subunit numbers and subunit sizes of enzymes: A study of the product of 100 human gene loci. *Annals of Human Genetics (London)*, **39**, 383–411.

ISHIMOTO, G., and KUWATA, M., 1974, Red cell glutamic-pyruvic transaminase polymorphism in Japanese populations. *Japanese Journal of Human Genetics*, **18**, 373–377.

JOHNSON, G., 1977, Isozymes, allozymes and enzyme polymorphisms: structural constraints on polymorphic variation. In *Isozymes II. Current Topics in Biological and Medical Research* (New York: Alan R. Liss, Inc.) pp. 11–19.

JOHNSON, G., 1979, Genetically controlled variation in the shapes of enzymes. *Progress in Nucleic Acid Research and Molecular Biology*, **22**, 293–326.

KOEHN, R. K., and EANES, W. F., 1977, Subunit size and genetic variation of enzymes in natural populations of *Drosophila*. *Theoretical Population Biology*, **11**, 330–341.

KOEHN, R. K., and EANES, W. F., 1978, Molecular structure and protein variation within and among populations. *Evolutionary Biology*, **10**, 39–100.

KIMURA, M., 1979, Model of effectively neutral mutations in which selective constraint is incorporated. *Proceedings of the National Academy of Sciences of the United States of America*, **76**, 3440–3444.

LI, W. H., 1978, Maintenance of genetic variability under the joint effect of mutation, selection and random drift. *Genetics*, **90**, 349–382.

LI, W. H., 1979 a, Maintenance of genetic variability under the pressure of neutral and deleterious mutations in a finite population. *Genetics*, **92**, 647–667.

LI, W. H., 1979 b, Effect of changes in population size on the correlation between mutation rate and heterozygosity. *Journal of Molecular Evolution*, **12**, 319–329.

MORGAN, K., and STROBECK, C., 1979, Is intragenic recombination a factor in the maintenance of genetic variation in natural populations? *Nature*, **277**, 383–384.

NEEL, J. V., 1973, Private genetic variants and the frequency of mutation among South American Indians. *Proceedings of the National Academy of Sciences of the United States of America*, **270**, 3311–3315.

NEEL, J. V., 1977, Some trends in the study of spontaneous and induced mutation in man. In *Human Genetics*, edited by S. Armendares and R. Lisker (Amsterdam: Excerpta Medica), pp. 19–32.

NEEL, J. V., 1978, Rare variants, private polymorphisms and locus heterozygosity in Amerindian populations. *American Journal of Human Genetics*, **30**, 465–490.

NEEL, J. V., and ROTHMAN, E. D., 1978, Indirect estimates of mutation rates in tribal Amerindians. *Proceedings of the National Academy of Sciences of the United States of America*, **76**, 1924–1928.

NEEL, J. V., SATOH, C., HAMILTON, H. B., OTAKI, M., GORIKI, K., KAGIOKA, T., FUJITA, M., NERISHI, S., and ASAKAWA, J., 1980, Search for mutations affecting protein structure in children of atomic bomb survivors: preliminary report. *Proceedings of the National Academy of Sciences of the United States of America*, **77**, 4221–4225.

NEEL, J. V., UEDA, N., SATOH, C., FERRELL, R. E., TANIS, R. J., and HAMILTON, H. B., 1978, The frequency in Japanese of genetic variants of 22 proteins. V. Summary and comparison with data on Caucasians from British Isles. *Annals of Human Genetics (London)*, **41**, 429–441.

NEI, M., 1976, Comments on "The intensity of selection for electrophoretic variants in natural populations of *Drosophila*", by B. D. G. Latter. In *Population Genetics and Ecology*, edited by S. Karlin and E. Nevo (New York: Academic Press), p. 409.

NEI, M., 1977, Estimation of mutation rates from rare protein variants. *American Journal of Human Genetics*, **29**, 225–232.

NEI, M., and CHAKRABORTY, R., 1976, Electrophoretically silent alleles in a finite population. *Journal of Molecular Evolution*, **8**, 381–385.

NEI, M., CHAKRABORTY, R., and FUERST, P. A., 1976, Infinite allele model with varying mutation rate. *Proceedings of the National Academy of Sciences of the United States of America*, **73**, 4164–4168.

NEI, M., FUERST, P. A., and CHAKRABORTY, R., 1978, Subunit molecular weight and genetic variability of proteins in natural populations. *Proceedings of the National Academy of Sciences of the United States of America*, **75**, 3359–3362.

NEI, M., and LI, W. H., 1976, The transient distribution of allele frequencies under mutation pressure. *Genetical Research (Cambridge)* **28**, 205–214.

OHTA, T., 1976, Role of very slightly deleterious mutations in molecular evolution and polymorphism. *Theoretical Population Biology*, **10**, 254–275.

ROTHMAN, E. D., and ADAMS, J., 1978, Estimation of expected number of rare alleles of a locus and calculation of mutation rates. *Proceedings of the National Academy of Sciences of the United States of America*, **75**, 5094–5098.

Schull, W. J., Ferrell, R. E., and Rothhammer, F., 1978, Genes, enzymes and hypoxia. In *Ecological Genetics: The Interface*, edited by P. Brussard (New York: Springer), pp. 73–90.

Sommer, S. S., and Cohn, J. E., 1980, The size distribution of proteins, mRNA and nuclear RNA. *Journal of Molecular Evolution*, **15**, 37–57.

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

Tchen, P., Ségur, J., Bois, E., Grenand, F., Fribourg-Blanc, A., and Fingold, N., 1978, A genetic study of two French Guiana Amerindian populations II: rare electrophoretic variants. *Human Genetics*, **45**, 317–326.

Turner, J. R. G., Johnson, M. S., and Evans, W. F., 1979, Contrasted modes of evolution in the same genome: Allozymes and adaptive changes in *Helicomius*. *Proceedings of the National Academy of Sciences of the United States of America*, **76**, 1924–1928.

Voelker, R. A., Schaffer, H. E., and Mukai, T., 1980, Spontaneous allozyme mutations in *Drosophila melanogaster*: rate of occurrence and nature of the mutants. *Genetics*, **94**, 961–968.

Ward, R. D., 1977, Relationship between enzyme heterozygosity and quaternary structure. *Biochemical Genetics*, **15**, 123–135.

Ward, R. D., 1978, Subunit size of enzymes and genetic heterozygosity in vertebrates. *Biochemical Genetics*, **16**, 799–810.

Welch, S. G., Mills, P. R., and Gaensslen, R. E., 1975, Phenotypic distributions of red cell glutamate pyruvate transaminase (E.C. 2.6.1.2.) isoenzymes in British and New York populations. *Human Genetics*, **27**, 59–62.

Wurm, S. A., 1975, Language distribution in the New Guinea area. In, *New Guinea area Languages and Language Study*, Vol. I, edited by S. A. Würm (Canberra, The Australian National University) Ser. C., No. 38, Pacific Linguistics, pp. 3–38.

Yasuda, N., 1973, An average mutation rate in man. *Japanese Journal of Human Genetics*, **18**, 279–287.

Zouros, E., 1979, Mutation rates, population sizes and amounts of electrophoretic variation of enzyme loci in natural populations. *Genetics*, **92**, 623–646.

Address correspondence to: K. K. Bhatia, Department of Human Biology, John Curtin School of Medical Research, P.O. Box 334, Canberra City, A.C.T., 2601, Australia.

Zusammtenfassung. Bisherige Untersuchungen von menschlichen Bevölkerungen fanden keine signifikante Ähnlichkeit zwischen genetischer Variabilität, gemessen durch die gesamte Heterozygotie, und der Cistron-Größe, gemessen durch das Molekulargewicht von Untereinheiten von Proteinen; es wurde jedoch gezeigt, daß die Zahl verschiedener seltener Allele in menschlichen Bevölkerungen mit der Größe von Untereinheiten korreliert. Die vorliegende Arbeit prüft diese Verknüpfungen weiter, wobei Daten über elektrophoretische Varianten auf 27 Loci bei 12 menschlichen Bevölkerungen mit einer Gesamtzahl von 800 000 Einzelbeobachtungen in Systemen benutzt werden.

Die Ergebnisse weisen darauf hin, daß bei Messung der genetischen Variabilität durch Heterozygotie seltener Allele anstelle der gesamten Heterozygotie eine signifkante Korrelation mit der Untereinheitsgröße vorhanden ist. Zusätzlich gibt es signifikante Unterschiede für die Heterozygotie seltener Allele zwischen multimer und monomer Proteinen, wobei die Breite der Variabilität bei Multimeren (und bei der Gesamtheit) geringer ist als bei Monomeren.

Schließlich hat die Heterozygotie seltener Allele eine erheblich größere Breite der Variabilität als die Untereinheitengröße. Im Gegensatz dazu ist die Breite der Heterozygotie seltener Allele zwischen den Bevölkerungen weniger als zehnfach, ein Faktor, der bei wirklicher Bevölkerungsgröße nicht erkennbar ist. Die Schätzungen der ralativen elektromorphen Mutationsraten (REMR) zwischen den Loci und auch zwischen den Bevölkerungen wurden berechnet, wobei sowohl die Verteilungen der Anzahl der verschiedenen seltenen Allele als auch die Heterozygotie seltener Allele benutzt wurden. Die Variationsbreite dieser Schätzungen ist wesentlich niedriger als die Schätzungen von Zouros (1979), der die gesamte Heterozygotie eingab.

Résumé. Les études précédentes de populations humaines n'ont pas réussi à trouver une relation significative entre la variabilité génétique, telle que mesurée par l'hétérozygotie totale, et le format du cistron, tel que mesuré par le poids moléculaire de la sous-unité de protéines, mais le nombre d'allèles rares différents dans les populations humaines s'est montré corrélé au format de la sous-unité. Le présent travail approfondit l'étude de ces relations, en utilisant des données sur des variantes électrophorétiques à 27 locus dans 12 populations humaines avec un total de 800 000 observations individuelles de systèmes.

Les résultats indiquent que, lorsque la variabilité génétique est mesurée par l'hétérozygotie des allèles rares au lieu de l'hétérozygotie totale, il y a une corrélation significative avec le format de la sous-unité. De plus, il y a des différences significatives pour l'hétérozygotie des allèles rares entre les protéines multimériques et monomériques, l'amplitude de la variabilité étant moindre pour les multimères (et au total) que pour les monomères.

Finalement, l'hétérozygotie des allèles rares a une amplitude de variabilité beaucoup plus large que celle du format des sous-unités. Par contre, l'amplitude de l'hétérozygotie des allèles rares entre populations est inférieure au décuple, un facteur qui ne ressort pas des volumes effectifs des populations. Les estimations interlocus comme interpopulationelles des taux relatifs de mutation électromorphe (REMR) ont été calculées, en employant les distributions du nombre de différents allèles rares aussi bien que l'hétérozygotie d'allèles rares. L'amplitude de ces estimations est très inférieure aux estimations données par Zouros (1979) employant l'hétérozygotie comme information.

Acta Anthropogenetica, Vol. 5 (1) : 67-87 (1981)

The Frequency of Private Electrophoretic Variants and Indirect Estimates of Mutation rate in Scheduled Tribes from South India

K.K. BHATIA

Department of Human Biology, John Curtin School of Medical Research Canberra, Australia

Key words

Electromorphs. Private Variants. Mutation Rate. Scheduled Tribes. South India.

Abstract

Data on private electrophoretic variants for 18 Scheduled Tribe populations from south India have been utilized to estimate mutation rate by two indirect procedures. The values of $\hat{\mu}$ for the total pooled data are 0.150×10^{-6} and 0.264×10^{-6}/locus per generation by the methods of Kimura and Ohta[30] and Nei[44] respectively. Three different groups of these tribes yield the unweighted average values of $\hat{\mu}$ as 0.193×10^{-6} and 0.410×10^{-6}/locus per generation by the two methods given above. The estimates on individual populations, however, show a wide variability, even if only the non-zero results are considered. The unweighted average of these individual tribe estimates is an order of magnitude higher than the estimates obtained for the total populations of all the 18 tribes.

The problems involved in estimating mutation rate from protein data using indirect methods in tribal populations of India are considerable because of their levels of detribalization and acculturation. The validity of the low values of $\hat{\mu}$ in these tribes, in comparison with the much higher estimates for the populations from the other parts of the world, is discussed.

Introduction

Recently Chakraborty and Roychoudhury[14] have published indirect estimates of mutation rate from protein data on some Scheduled Tribes from south India. Their estimates differ from other such estimates by more than an order of magnitude, generated on Amerindians[38 42 44 61], Australian Aborigines[6 8] and Papua New Guineans[9]. Large errors, however, are known to be associated with estimates derived by using the three available procedures of Kimura and Ohta[30], Nei[44] and Rothman and

Adams[56]. Chakraborty and Roychoudhury[14] have discussed some of these problems with special reference to the data used by them for moderately acculturated and demographically expanding south Indian tribal groups.

In the last few years this department has screened a number of Scheduled Tribes for genetic polymorphisms over subsets of 12 to 23 loci. The populations studied are : Kadars[57], Todas, Kurumbas, Irulas and Malayaryans[58], Kotas[24-26] Savaras and Jatapus[52] Kolams[51], Chenchus[50], Raj Gonds, Pardhans, Koyas, Konda Reddis, Lambadis and Yerukulas[11], Konda Kammaras, Koyas (second series) and Gadabas (unpublished material).

A number of other laboratories have reported on red cell and serum proteins in Andhra Pradesh tribals[5 18 19 27 28 48 53-55 59 60]. In addition comparative results are available for the non-tribal populations of south India (see Basu, 1978 for a list of references) as well as tribal and non-tribal populations from the adjoining states of Maharashtra, Madhya Pradesh and Orissa. The latter information is valuable in defining variants.

Data are now available for 18 tribal populations of south India sampled from a total set of 30 protein loci. Using this information mutation rate estimates for tribes from India have been generated which supplement the results given by Chakraborty and Roychoudhury[14].

The study population

The 18 tribal populations included in this study have been divided into three groups on the basis of their geographical proximity and demographic features. *Group I* comprises nine tribal populations from the northern (Adilabad, Warrangal, Khammam, Srikakulam, Vishakhapatnam, E. Godavari and W. Godavari) districts of Andhra Pradesh; these populations have been grouped together because of their relatively large population sizes, continuous dispersion, positive growth trends in the past 100 years and cultural affiliations with the Scheduled Tribes of central India. In some cases data from the same, or adjoining districts have been pooled. *Group II* includes Chenchus, Lambadis and Yerukulas from Mahabubnagar and Kurnool districts in southern Andhra Pradesh. They have been grouped together because they were sampled for the same districts but have a discontinuous distribution in restricted pockets over large areas with small effective breeding units. *Group III* is constituted by six small tribes, all restricted to the Nilgiris and Annamalai Hills of Kerala and Tamil Nadu States. The list of tribes studied is given in Table I.

Only populations screened for at least five protein loci have been included in the survey. Table II gives the number of persons tested for the

total 30 red cell and serum protein loci screened. Data generated by the use of non-electrophoretic methods for haemoglobins and glucose-6-phosphate dehydrogenase have not been utilized.

Number of private and rare allelic variants

An allelic variant is considered to be 'private' if it occurs uniquely in only one population[39]. In addition, if any variant allele has a frequency of less than 1% it is considered to be 'rare'. If an electrophoretic variant has been reported in populations from a number of localities in the subcontinent (e g. PHI 2-1), the corresponding allele has not been included in the list of rare variants since its presence in any particular population could be due to intermixture.

In Table III the private and rare variants are indicated in addition to rare variants which are found in more than one population. The latter include especially $Hb\beta^{S}$, $LDH_{B}{}^{Cal-1}$ and PGD^{C} which are present at low frequency or are absent in some of these tribes. A few others (e.g. PHI^{3}), though rare in general, sometimes achieve polymorphic frequency in one particular population. All of the variants in this latter category, of course, are excluded from the calculation of mutation frequency.

Group I

Seven private allelic variants are restricted to the populations of this group (Table III). Two of these assume polymorphic proportions. PGD^{Gadaba} (1.14%) in Gadabas (unpublished data) and $Hb\alpha^{Koya\ Dora}$ (5.00%) in a sample of Koyas from Polavaram Taluk in W. Godavari district[19].

Group II

Only two private allelic variants, both polymorphic, were detected in group II populations, PHI^{5} and PGM_{2}^{10}. PHI^{5} has been included although a single copy of PHI^{5} has been reported previously from north India[10]. It is unlikely that its presence in the Chenchus is due to admixture.

Group III

A number of private (rare as well as polymorphic) variants have been observed in three of the six tribal populations in this group. The private polymorphisms are PGM_{1}^{6K}, PGD^{Kadar} and PGM_{1}^{6Mal}, the former two in Kadars and the latter in Malayaryans, and two rare variants *Pep* B^{K} in Kadars and $LDH_{A}{}^{Toda}$ in Todas.

Table I. Actual population size (N) in Scheduled Tribes of south India

Sl. No.	Scheduled Tribe	District/State of enumeration	Density per km² (D)	Total census size	Proportion in 15-44 yrs age group (λ)	Actual population size (N)
	*Group I***					
1.	Savaras	Srikakulam[1]	8.00	40,228	0.439	9,899
2.	Jatapus	Srikakulam[1]	7.61	38,250	0.461	9,811
3.	Kolams	Adilabad[1]	1.62	8,150	0.402	1,836
4.	Koyas	Adilabad[1]	11.56	58,140	0.408	13,297
5.	Raj Gonds	Adilabad[1]	7.21	36,275	0.400	8,136
6.	Pardhans	Yeotmal[2]	2.42	12,171	0.441	3,003
7.	Konda Reddis	E. Godavari[1]	3.45	17,333	0.407	3,954
8.	Konda Kammaras	Vishakhapatnam[1]	1.12	5,619	0.432	1,360
9.	Gadabas	Srikakulam[1] Vishakhapatnam†	1.03	5,201	0.434	1,264
	*Group II**					
10.	Chenchus	Mahabubnagar & Kurnool[1]	—	7,984	0.403	3,217
11.	Lambadis	Kurnool[1]	—	11,704	0.385	4,511
12.	Yerukulas	Kurnool[1]	—	10,650	0.409	4,357
	Group III					
13.	Todas	Kerala	—	930	0.447	416
14.	Kurumbas	Kerala/Tamil Nadu	—	4,073	0.506	2,063
15.	Irulas	Kerala/Tamil Nadu	—	103,039	0.445	45,972
16.	Malayaryans	Kerala	—	4,194	0.556	2,331
17.	Kotas	Tamil Nadu	—	1,188	0.454	539
18.	Kadars	Kerala/Tamil Nadu	—	1,418	0.481	681

* The estimates are adjusted to 1921 level.
** In Group I, census size is given as the 'size of the neighbourhood'; for explanations see text.
1 Andhra Pradesh
2 Maharashtra
† E. Godavari, W. Godavari and Warrangal Districts.

Table II. List of red cell enzymes, proteins and serum proteins included in the study

Sl. No.	System	Abbreviation	Sample size	Groups (n>1000)	Populations studied*
	Red Cell Enzymes				
1.	Acid phosphatase—1	ACP_1	3,619	I, II	1-18
2.	Adenosine deaminase	ADA	60	—	1
3.	Adenylate kinase—1	AK_1	3,824	I, II	1-7, 9-18
4.	Carbonic anhydrase—1	CA_1	394	—	9, 10, 12
5.	Carbonic anhydrase—2	CA_2	1,021	—	3, 9, 10, 12, 17
6.	Esterase D	EsD	2,023	I	3, 4, 8-10, 12, 17
7.	Glucose-6-phosphate dehydrogenase	G-6-PD	213	—	18
8.	Glutamic oxaloacetic acid transaminase	GOT	359	—	1-3
9.	Glyoxalase—1	GLO_1	756	—	4, 8, 17
10.	Isocitrate dehydrogenase	ICD	1,497	II	3, 10, 12-18
11.	Lactate dehydrogenase—A	LDH_A	3,575	I, II	1-18
12.	Lactate dehydrogenase—B	LDH_B	3,575	I, II	1-18
13.	Malate dehydrogenase—2	MDH_2	3,646	I, II	1-18
14.	Nucleoside phosphorylase	Np	580	—	3, 13-16, 18
15.	Peptidase A	Pep A	1,012	—	1-3, 13-16, 18
16.	Peptidase B	Pep B	1,012	—	1-3, 13-16, 18

Table II. Contd.

Sl. No.	System	Abbreviation	Sample size	Groups (n>1000)	Populations studied*
17.	Peptidase D	Pep D	693	—	1-3, 13-16
18.	Phosphogluconate dehydrogenase	PGD	3,610	I, II	1-18
19.	Phosphoglycerate kinase	PGK	2,804	I, II	1-3, 9, 10, 12-18
20.	Phosphoglucomutase—1	PGM_1	4,014	I, II	1-18
21.	Phosphoglucomutase—2	PGM_2	4,022	I, II	1-18
22.	Phosphohexose isomerase	PHI	3,644	I, II	1-18
23.	Superoxide dismutase	SOD_A	3,644	I, II	1-18
	Red Cell Proteins				
24.	Hemoglobin—α	Hb-α	4,159	I, II	1-18
25.	Hemoglobin—β	Hb-β	4,159	I, II	1-18
	Serum Proteins				
26.	Albumin	Alb	1,829	—	4-6, 10, 11, 13-16, 18
27.	Caeruloplasmin	Cp	2,208	II	4-6, 11, 13-18
28.	Group specific component	Gc	1,006	—	4-6, 11, 14, 15
29.	Haptoglobin	Hp	2,477	II	4-6, 10, 11, 13-18
30.	Transferrin	Tf	2,374	II	4-6, 10, 11, 13-18

* The serial numbers of the populations are given in Table I.

Table III. Rare variants and private polymorphisms in Scheduled Tribes of South India.

Population	Single locus determinations	Allelic variants (percent frequency)	References
Group I			
Savaras	2,150	$Hb\beta^{s}$ (0.76)	52
Jatapus	2,664	$Hb\beta^{s}$ (0.64), PHI^{3} (0.65)	52
Kolams	3,818	PHI^{2} (0.23), PGM_1^{7} (0.47)	51
Koyas	7,897	$Hb\alpha^{\text{Koya Dora}}$ (0.99)*, $Hb\alpha^{\text{Rampa}}$ (0.05)*, PHI^{2} (0.19), PHI^{3} (1.36), $Alb^{\text{Koya Dora}}$ (0.09)*, $Tf^{\text{D Chi}}$ (2.92), PGM_1^{4} (0.11)*, PGM_1^{6} (0.11)	5, 11, 18, 19, 28, 53, 54, 59, 60, Unpublished
Raj Gonds	3,559	PHI^{3} (0.37), PGD^{C} (0.75), $Tf^{\text{D Gond}}$ (0.70)*, $Tf^{\text{D Chi}}$ (1.04)	11, 28, 53, 54
Pardhans	1,952	—	11, 28, 53, 54
Konda Reddis	1,657	ACP_1^{C} (0.55)	11, 18, 19, 59, 60
Konda Kammaras	1,429	PHI^{2} (0.45), PGD^{C} (0.93)	Unpublished
Gadabas	13,035	PHI^{6} (0.45)*, PGD^{Gadaba} (1.14)*, PGM_1^{7} (0.10)	Unpublished

Table III. Contd.

Population	Single locus determi-nations	Allelic variants (percent frequency)	References
Group II			
Chenchus	3,521	$Hb\beta^{s}$ (0.26), PHI^{2} (0.25), PHI^{5} (2.71)*, LDH_{B}^{Cal-1} (1.48) Tf^{D} (0.35), PGM_{1}^{1} (0.25)	50
Lambadis	1,674	$Tf^{D\ Chl}$ (0.32)	11, 28, 53, 54
Yerukulas	569	PGM_{1}^{1} (7.69), $PGM_{2}^{1\,0}$ (1.25)*	11
Group III			
Todas	2,303	$Hb\beta^{s}$ (0.51), PGD^{C} (0.51), LDH_{A}^{Toda} (0.51)*	32, 58
Kurumbas	1,049	LDH_{B}^{Cal-1} (3.49)	31, 32, 58
Irulas	3,765	LDH_{B}^{Cal-1} (0.29)	31, 32, 58
Malayaryans	1,280	$PGM_{1}^{6\,Mal}$ (7.63)*	58
Kotas	10,816	—	24, 25, 26
Kadars	4,671	$Hb\beta^{s}$ (0.47), LDH_{B}^{Cal-1} (1.64), PGD^{K} (4.24)* ACP_{1}^{C} (0.48), PGM_{1}^{6K} (2.11)*, *Pep* B^{K} (0.23)*	57

*Private (rare/polymorphic) allele.

Table IV. Indirect estimates of mutation rate in south Indian Scheduled Tribes

Serial number	Population	Number of cistrons	Mean per locus determinations (n_I)	Private rare variants	$\hat{t}_o$**	$\mu \times 10^{-6}$	
						K–O method	Nei's method
	*Group I**						
1.	Savaras	18	179.44 ± 6.31	0(0)	16.205	—	—
2.	Jatapus	17	156.71 ± 0.19	0(0)	16.202	—	—
3.	Kolams	21	181.61 ± 10.21	0(0)	13.446	—	—
4.	Koyas	19	415.63 ± 57.75	4(4)	16.689	0.47 ± 0.28	1.86 ± 1.09
5.	Raj Gonds	17	209.75 ± 25.65	1(1)	15.884	0.23 ± 0.23	1.26 ± 1.26
6.	Pardhans	17	114.82 ± 9.98	0(0)	14.251	—	—
7.	Konda Reddis	12	138.08 ± 24.12	0(0)	14.702	—	—
8.	Konda Kammaras	13	109.92 ± 0.08	0(0)	12.954	—	—
9.	Gadabas	16	814.69 ± 86.13	2(1)	12.834	3.85 ± 2.57	4.43 ± 4.43
	AVERAGE				14.796± 0.501	0.51 ± 0.42	0.84 ± 0.51
	*Group II**						
10.	Chenchus	20	176.05 ± 6.84	1(0)	14.364	0.54 ± 0.54	—
11.	Lambadis	17	98.47 ± 14.11	0(0)	14.918	—	—
12.	Yerukulas	17	33.47 ± 2.42	1(0)	14.861	0.45 ± 0.45	—
	AVERAGE				14.714± 0.176	0.33 ± 0.17	—

Table IV. Contd.

Serial number	Population	Number of cistrons	Mean per locus determinations (n_I)	Private rate variants	$\hat{t}_o$**	$\mu \times 10^{-6}$ K—O method	$\mu \times 10^{-6}$ Nei's method
	Group III						
13.	Todas	22	104.68 ± 5.26	1(1)	8.741	6.21 ± 6.21	36.97 ± 36.97
14.	Kurumbas	23	45.61 ± 2.81	0(0)	10.823	—	—
15.	Irulas	23	163.70 ± 8.26	0(0)	14.858	—	—
16.	Malayaryans	22	58.18 ± 0.52	1(0)	10.981	0.88 ± 0.88	—
17.	Kotas	20	540.80 ± 6.93	0(0)	9.078	—	—
18.	Kadars	22	212.32 ± 0.36	3(1)	9.382	10.61 ± 5.83	11.53 ± 11.53
	AVERAGE				10.644± 0.923	2.95 ± 1.82	8.08 ± 6.08
	OVERALL AVERAGE (18 tribes)					1.29 ± 0.67	3.11 ± 2.10

* Actual population size adjusted to 1921 census numbers.

** The value of $N_e v/N$ is 0.819, 0.819 and 0.650 in Group I, Group II and Group III respectively.

Table V. Indirect estimates of mutation rate in three major groups of south Indian Scheduled Tribes

Population group	Number of loci	Single locus determi-nations	Private (rare) variants	N (15-44 yrs)	$\mu \times 10^{-6}$	
					μK—O	μNEI
			ALL LOCI			
Group I	29	38,161	7 (7)	52,560	0.156±0.060	0.351±0.136
Group II	22	5,764	2 (1)	12,085	0.256±0.177	0.568±0.568
Group III	27	23,884	5 (5)	52,002	0.167±0.084	0.310±0.156
Average					0.193±0.032	0.410±0.080
			LOCI WITH n > 1,000 TESTS			
Group I	14	30,868	5 (5)	52,560	0.230±0.109	0.449±0.213
Group III	17	19,435	4 (4)	52,002	0.213±0.123	0.361±0.209
Average					0.222	0.605
			TOTAL POOLED DATA			
Total loci	30	67,809	14 (14)	116,089	0.150±0.048	0.264±0 084
Loci with n>1,000	23	64,754	14 (14)	116,089	0.196±0.059	0.325±0.099

Methodology

Three methods for estimating the mutation rate indirectly from electromorphs have been suggested so far : Kimura and Ohta[30], Nei[44] and Rothman and Adams[56]. These Methods have been summarised by Bhatia *et al*[8]. In the present paper results have been included only for the first two of these methods. Because of small mean sample sizes and the low recovery of rare variants the estimates based on Rothman and Adams' method in the present populations are highly inflated.

In both the Kimura and Ohta's and Nei's methods the estimation of actual (apparent) population size N, (effective population size in Nei's[44] and Bhatia *et al*[8] terminology) is an important parameter and is normally equated to the number of breeding individuals given as the proportion of the population in the age group 15-44 years. In populations with cyclic changes in population size over the past few generations, Wright[68] has recommended the use of harmonic mean.

To accommodate the role of isolation by distance in the large, continuously dispersed populations of group I, the value of N is estimated as the 'size of neighbourhood', following Wright[69], as

$$N = 4\pi\sigma^2 D\lambda$$

where σ^2 is the variance of migrational distances, D is the population density and λ the proportion of the reproductive age group (15—44 years) in the population. Pingle's[49] data yield the variance of marital distance in Adilabad tribes to be approximately 400km^2. Majumdar[33] has, however, given much smaller values for marital distances for the Andhra populations as a whole. The individual values of D and λ estimated from age and sex tables of the Andhra tribes[12] are shown in Table I.

The effective breeding unit in the discontinuously distributed populations of group II is taken to be the administrative district, where the samples have been collected. Total census sizes have been utilized in Group III populations for estimating N.

Except Chenchus, who have increased marginally over their 1911 numbers, the populations of groups I and II have been adjusted for population increase since 1881. Taking 1921 census as the base level, which is quite close to the harmonic mean size since 1881, we calculate the new estimates of N to be 0.560 of their 1971 numbers. No such adjustment is necessary for group III tribes, which have only recently built up their original numbers to 1881 levels after a decline in the early decades of this century.

Another important parameter used in estimating mutation rates by Kimura and Ohta's method is the expected number of generations a

mutant survives prior to extinction $\bar{t}_0$. The estimate is given as

$$\hat{\bar{t}}_o = 2\left[\frac{N_{ev}}{N}\right] \ln [2N]$$

where N is the actual (apparent) population size as defined earlier and N_{ev} is the variance effective number. The estimation of $\bar{t}_o$, however, involves large standard errors[30].

The value of the variance effective number N_{ev}, is given[16 17] as

$$N_{ev} = 2N/[(1-F)+(1+F)\ V_k/\bar{k}]$$

where F is a measure of departure from Hardy Weinberg proportions, taken formally equivalent to the inbreeding coefficient and $\bar{k}$ and V_k are mean and variance respectively of the progeny size surviving to adulthood. $V_k/\bar{k}$ also defines the index of variability. At birth and adulthood the mean and variance will be defined by $\bar{k}_b$ and V_{kb}, $\bar{k}_a$ and V_{ka} respectively.

Murty and Ramesh[37] and Ghosh[21] have provided the estimates of $\bar{k}_b$ and V_{kb} for post-reproductive age women and also the index of mortality I_m[15], for Adilabad tribes and Kotas respectively. The index of variability at adulthood ($V_{ka}/\bar{k}_a$) is recalculated using the formulae

$$P_s = 1/(1+I_m)$$

$$\text{and } \frac{V_{ka}}{\bar{k}_a} = 1 + P_s\left[\frac{V_{ka}}{\bar{k}_b} - 1\right]$$

where P_s is the probability of survival to adulthood and the subscripts a and b refer to the values at adulthood and at birth respectively. The estimated values of this index in Adilabad tribes and Kotas is 1.43 and 1.95 respectively. Basu's[2] data yield a value of 2.03 for Kotas. The high value in Kotas is attributed to a large proportion of nulliparous women in the 45 + years age group. Similar demographic trends are seen in Irulas[4]. Since no published results are available on progeny size for men and women separately, adjustments for variation due to polygamy are not made in these calculations.

The value of F obtained from pedigree data on Andhra tribes and Kotas is 0.030[65] and 0.040[22 23] respectively. Inserting these values of F and the respective estimates of $V_{ka}/\bar{k}_a$ in the equation, N_{ev}/N becomes 0.819 and 0.650 in Andhra tribes and Kotas respectively. The former value is used for computing $\bar{t}_0$ in individual populations of groups I and II, and the latter for the populations of group III. The estimates $\hat{\bar{t}}_o$, are given in Table IV.

Estimates of mutation rates

The results on the indirect estimates of mutation rate obtained at three levels of population organisation, i. e., at individual population, individual group and all tribes level are presented in Tables IV and V. The estimators used are $\hat{\mu}_{K-O}$ and $\hat{\mu}_{NEI}$, as given by Kimura and Ohta[30] and Nei[44] respectively.

At individual population level the estimates of μ show wide variability (Table IV). Even for non-null results the values differ by more than an order of magnitude.

The estimates of μ in group III populations are on an average higher than those obtained for groups I and II populations. The unweighted average of these 18 individual population estimates is $1.29\times10^{-6}\pm0.67\times10^{-6}$/locus per generation and $3.11\times10^{-6}\pm2.10\times10^{-6}$/locus per generation by the methods of Kimura and Ohta[30] and Nei[44] respectively. These estimates, however, entail large standard errors (SE) which may be contributed by fluctuations in these estimates of $\hat{I}$ and $\hat{I}_q$ as also the errors associated with the estimation of N.

The estimates $\hat{\mu}_{K-O}$ and $\hat{\mu}_{NEI}$ at individual group level, however, do not show much variability. Theu nweighted averages of three individual group estimates are $0.193\times10^{-6}\pm0.032\times10^{-6}$ and $0.410\times10^{-6}\pm0.080\times10^{-6}$/ locus per generation by the procedures of Kimura and Ohta[30] and Nei[44] respectively. The standard errors (SE) of the individual group estimates, obtained from the number of rare alleles over the surveyed loci are not so large, except for the estimates on group II (Table V).

The pooled data, over all the 18 populations, yield much smaller estimates of μ. The values of $\hat{\mu}_{K-O}$ and $\hat{\mu}_{NEI}$ are $0.150\times10^{-6}\pm0.048\times10^{-6}$/ locus per generation and $0.264\times10^{-6}\pm0.084\times10^{-6}$/locus per generation respectively. The terms of standard errors (SE) given are due to the variance of $\hat{I}$ and $\hat{I}_q$, estimated as 0.649.

The large standard errors (SE) associated without the estimates of I and I_q are largely due to fluctuations in the sample size which affects the recovery of rare alleles seriously and the variability in the mutation rate over loci on account of subunit size (MW) variations. The variability in the sample size of loci tested for group I is 60—2,589, for group II is 113—1,327 and for the total pooled data is 60—4,159. Since some loci were tested on a relatively small number of individuals we have estimated new values of μ only for those loci which have been tested for at least 1,000 individuals. The new estimates of μ_{K-O} for groups I and II are $0.230\times10^{-6}\pm0.109\times$

10^{-6}/locus per generation and $0.213\times10^{-6}\pm0.123\times10^{-6}$/locus per generation respectively with a mean value of 0.222×10^{-6}/locus per generation. Similar estimates of μ_{NEI} are shown in Table V. These estimates are slightly higher than our earlier estimates for groups I and II by the two methods.

The estimates of μ_{K-O} and μ_{NEI} for pooled data over 23 loci ($n>1,000$) are $0.196\times10^{-6}\pm0.059\times10^{-6}$/ locus per generation and $0.325\times10^{-6}\pm 0.099\times10^{-6}$/locus per generation (Table V). The differences of these estimates from those obtained previously are only marginal.

Discussion

The indirect estimates of mutation rate generated on the Scheduled Tribes from south India are clearly outside the range of similar estimates when comparisons are made with those obtained at similar levels of population organisation on Amerindians[42], Australian Aborigines[8] and Papua New Guineans[9]. At individual tribe level, the unweighted average of μ for tribes in India is an order of magnitude less than the unweighted average for Amerindians. Similarly, the results on the pooled population of all tribes from south India are considerably lower than similar estimates on the Australian Aborigines and the Papua New Guineans. The present results, although based on a much wider data base, in fact, confirm the apprehensions of Chakraborty and Roychoudhury[14] regarding the use of data on moderately acculturated Indian tribes for the estimation of mutation rate, although the possibility of regional/ethnic differences in mutation rate exists[43].

One of the factors which affect these estimates on Indian tribes seriously are the conservative procedures employed in designating a private variant. In the Indian context, where a large number of communities live together in the same area, identities between electromorphs suggest common descent and fresh mutations are private by default only. Considerable under-estimation of this sort of data makes the genetic interpretation of these results difficult.

Another serious source of error in utilising the indirect procedures is the use of the parameter N, the actual size of the population. In continuously distributed, large sized population groups the effective size of the breeding unit estimated by the methods of Wright[63] or Bhatia *et al*[9], suitability of the approach notwithstanding, is only approximate and tends to err on the higher side. In the absence of hard data on the historic demography of these pre-literate societies, analogous dilemmas of a more temporal nature are faced. In addition, the extrapolation of the current demographic compositions to a time-specific constancy is disput-

able. The much lower estimates of μ, in the Papua New Guineans[9] and the Indian tribes (present study) may be attributed partly to these over-estimations of the expected harmonic values of N.

The use of electrophoretic data, as analysed by the standard methods, clearly defines only a subset of total mutational events occurring at a given cistron and thus any estimates obtained by these approaches must be adjusted for these under-estimations. In addition to about two thirds of the aminoacid substitutions which lead to no charge change[34 45 61], a large fraction, depending upon the distribution/density of the population and the relative frequencies of the electromorphs, of electrophoretically detectable substitutions is lost due to coalescence with other electromorphs[13 46 63]. The effect of the latter is correlated with the population size. For presumably similar neutral mutation rates over similar sets of protein loci, Bhatia, *et al*[9] found 2.66 times more silent alleles in the numerically stronger (and more densely distributed) Papua New Guinean communities, than in the thinly spread, small sized group of the Australian Aborigines. Because of the undefinable nature of these population sizes no adjustments have been made on this accord, though the present estimates may only be 20 to 40 per cent of the real values.

Another class of mutations omitted in these calculations is null mutations. Although the biochemical nature of these mutations may range from a single aminoacid substitution in a polypeptide to a total loss of polypeptide and, in theory, may result from mutations either in structural or regulatory genes[40], the ratio of μ_{null} to $\mu_{variant}$ is known to range from 2-6 fold[36 66]. Arthur *et al*[1] and Nelson and Harris[47] have reported more than 12 fold more null mutations in experiments on mutagenised human cultured cells. Since it is now possible to distinguish between structural and regulatory mutations[62] the proportion of null mutations at structural loci can be estimated. Although we do not introduce any correction for this factor here, it may be noted that such adjustments will raise the estimates considerably, especially on large sized populations.

The procedures for calculating indirect estimates of μ from protein data have now been extended to non-human species by McCommas and Chakraborty[35]. In *Bunodosoma Cavernata* they have estimated the $\hat{\mu}$ to be 6.3×10^{-7} to 6.3×10^{-8}/locus per generation for population sizes of 10^6 to 10^7 individuals. These results, along with our results of Indian and Papua New Guinean populations, indicate that very low values of $\hat{\mu}$ are generated from protein data by using indirect procedures, specially if the population sizes are large.

The results on direct estimates of mutation rates from protein data on

Drosophila and man are now available. Mukai and Cockerham[36] and Voelker *et al*[67] reported the frequency of band morph mutations in *Drosophila melanogaster* as 1.81×10^{-6} and 1.28×10^{-6}/locus per generation respectively. Dubinin and Altukhov[20] and Neel *et al*[43] have given these estimates in human populations as 6×10^{-5} and 0.34×10^{-5} locus per generation in Russians and Japanese respectively. The results on human populations are, however, difficult to evaluate since more than 522, 119 determinations in English[29], Amerindians[41] and Japanese[43] have failed to identify a single confirmed instance of spontaneous mutation, although the possibility of detecting much common null mutations also exists. If anything, these results only indicate that the differences in mutation rates between moderately acculturated, comparatively large sized, Scheduled Tribe groups of south India, and the other non-tribal communities, may not be very large. Bhatia[7] has also indicated that the range of inter-populational estimates of relative electromorph mutation rates is much lower than the range of their effective population sizes, indicating that mutability differences among human populations, both civilized and primitive, if any, are only marginal.

Acknowledgements

I am grateful to Dr R.L. Kirk for the invaluable suggestions in the preparation of this manuscript and Dr N.M. Blake for help in the compilation of data.

References

1. Arthur E, Steel CM, Evans HJ, Povey S, Watson B and Harris H (1975) Genetic studies on human lymphoblastoid cell lines. Isozymes and cytogenetic heterogeneity in a cell line with evidence for localization of Pep A locus in man. Ann Hum Genet *39* : 33-42.
2. Basu A (1972) A demographic study of the Kota of Nilgiri Hills. J Ind Anthrop Soc *7* : 29-45.
3. Basu A (1978) Physical anthropological research in south India : a bibliographical review. J Ind Anthrop Soc *18* : 187-213.
4. Basu MP (1967) A demographic profile of the Irula. Bull Anthrop Surv India *16* : 267-289.
5. Bernini LF, De Jong WW and Meera Khan P (1970) Varianti emglobiniche mella populazione tribale dell 'Andhra Pradesh. Molteplicita del locus α^{Hb} nell' uomo. Atti Assoc Genet Ital *15* : 191-194.
6. Bhatia K (1980) Factors affecting electromorph mutation rates in man : an analysis of data from Australian Aborigines. Ann Hum Biol *7* : 45-54.
7. Bhatia K (1981) Rare allele heterozygosity and relative electromorph mutation rates in man. Ann Hum Biol (Submitted).

8. Bhatia K, Blake, NM and Kirk, RL (1979) The frequency of private electrophoretic variants in Australian Aborigines and indirect estimates of mutation rate. Amer J Hum Genet *31* : 731-740.

9. Bhatia K, Blake NM, Serjeantson SW and Kirk, RL (1981) The frequency of private electrophoretic variants and indirect estimates of mutation rate in Papua New Guinea. Amer J Hum Genet (In press).

10. Blake NM, Kirk RL, McDermid, EM, Omoto K and Ahuja YR (1971) The distribution of serum protein and enzyme group systems among north Indians. Hum Hered *21* : 440-457.

11. Blake NM, Ramesh A, Vijayakumar M, Murty JS and Bhatia KK (1981) Genetic studies on some tribes of the Telangana region, Andhra Pradesh. Acta Anthropogenet 5 : 41-56.

12. Census of India Special tables on Scheduled Castes and Scheduled Tribes. Series 2 Andhra Pradesh, Part V-A (Directorate Census Operations Hyderabad 1971).

13. Chakraborty R and Nei M (1976) Hidden genetic variability within electromorphs in finite populations. Genetics *84* : 385-393.

14. Chakraborty R and Roychoudhury AK (1978) Mutation rates from rare variants of proteins in Indian tribes. Hum Genet *43* : 179-183.

15. Crow JF (1958) Some possibilities for measuring selection intensities in man. Hum Biol *30* : 1-13,

16. Crow JF and Kimura M (1972) The effective number of a population with overlapping generations : a correction and further discussion. Amer J Hum Genet *24* : 1-10.

17. Crow JF and Morton NE (1955) Measurement of gene frequency drift in small populations. Evolution *9* : 202-214.

18 De Jong WW, Bernini LF and Meera Khan P (1971) Haemoglobin Rampa: α 95 Pro→Ser. Biochim Biophys Acta *236* : 197-200.

19. De Jong WW, Meera Khan P and Bernini LF (1975) Haemoglobin Koya Dora: high frequency of a chain termination mutant. Amer J Hum Genet *27* : 81-90.

20. Dubinin NP and Altukhov Yu P (1979) Gene mutations (*de novo*) found in electrophoretic studies of blood proteins of infants with anomalous development. Proc Natl Acad Sci USA *76* : 5226-5229.

21. Ghosh AK (1970) Selection intensity in Kota of Nilgiri Hills, Madras. Soc Biol *17* : 224-225.

22. Ghosh AK (1972) Inbreeding in the Kota of Nilgiri Hills, Madras. Soc Biol *19*: 289-291.

23. Ghosh AK (1976) The Kota of the Nilgiri Hills : a demographic study. J Biosoc Sci *8* : 17-26.

24. Ghosh AK (1977) The distribution of genetic variants of glyoxalase I, esterase D and carbonic anhydrase I and II in Indian populations. Indian J Phys Anthrop Hum Genet *3* : 73-83.

25. Ghosh, AK (1977) Polymorphism of red cell glyoxalase I with reference to southeast Asia and Oceania. Hum Genet *39* : 91-95.

26. Ghosh AK, Kirk RL, Joshi SR and Bhatia HM (1977) A population genetic study of the Kota in the Nilgiri Hills, south India. Hum Hered *27* : 225-241.

27. Goud, JD and Rao PR (1977) Distribution of some genetic markers in the Yerukala tribe of Andhra Pradesh. J Ind Anthrop Soc *12* : 258-265.

28. Goud JD and Rao PR (1980) Transferrin, haptoglobin and group specific component types in tribal populations of Andhra Pradesh. Hum Hered *30* : 12-17.

29. Harris H, Hopkinson DA and Robson EB (1974) The incidence of rare alleles determining electrophoretic variants. Data on 43 enzyme loci in man. Ann Hum Genet *37* : 237-253.

30. Kimura M and Ohta T (1969) The average number of generations until extinction of an individual mutant gene in a finite population. Genetics *63* : 701-709.

31. Kirk RL, Cleve H and Bearn AG (1963) The distribution of the group specific component (Gc) in selected populations in South and S. E. Asia and Oceania. Acta Genet (Basel) 13 : 140-149.

32. Kirk RL, Lai LYC, Vos GH, Wickremsinghe RL and Perera DJB (1962) The blood and serum groups of selected populations in south India and Ceylon. Amer J phys Anthrop *20* : 485-497.

33. Majumdar PP (1977) Matrimonial migration : a review with special reference to India. J Biosoc Sci *9* : 381-401.

34. Marshall DR and Brown, AHD (1975) The charge state model of protein polymorphisms in natural populations. J Mol Evol *6* : 149-163.

35. McCommas SA and Chakraborty R (1980) Estimation of mutation rates in three species of sea anemone in the genus *Bundosoma*. Genetics *94* : s66.

36. Mukai T and Cockerham CC (1977) Spontaneous mutation rates at enzyme loci in *Drosophila melanogaster*. Proc Natl Acad Sci USA *74* : 2514-2517.

37. Murty JS and Ramesh A (1979) Selection intensities among the tribal populations of Adilabad District, Andhra Pradesh, India. Soc Biol *25* : 302-305.

38. Neel JV (1973) Private genetic variants and the frequency of mutation among South American Indians. Proc Natl Acad Sci USA *70* : 3311-3315.

39. Neel JV (1978) Rare variants, private polymorphisms and locus heterozygosity in Amerindian populations Amer J Hum Genet *30* : 465-470.

40. Neel JV (1978) Mutation and disease in man. Can J Genet Cytol *20* : 295-306.

41. Neel JV, Mohrenweiser HW and Meisler MH (1980) Rate of spontaneous mutation at human loci encoding protein structure. Proc Natl Acad Sci USA *77* : 6037-6041.

42. Neel JV and Rothman ED (1978) Indirect estimates of mutation rates in tribal Amerindians. Proc Natl Acad Sci USA *75* : 5585-5588.

43. Neel JV, Satoh C, Hamilton HB, Otake M, Goriki K, Kageoka T, Fujita M, Neriishi S and Asakawa J (1980) Search for mutations affecting protein structure in children of atomic bomb survivors : Preliminary report. Proc Natl Acad Sci USA *77* : 4221-4225.

44. Nei M (1977) Estimation of mutation rates from rare protein variants. Amer J Hum Genet *29* : 225-232.

45. Nei M and Chakraborty R (1973) Genetic distance and electrophoretic identity of proteins between taxa. J Mol Evol *2* : 323-328.

46. Nei M and Chakraborty R (1976) Electrophoretically silent alleles in a finite population. J Mol Evol *8* : 381-385.

47. Nelson RL and Harris H (1978) The detection of mutation in human diplo fibroblasts after mutagen treatment using non-selective cloning and enzyn electrophoresis. Mut Res *50* : 277-283.

48. Papiha SS, Bernal JE, Roberts DF, Habeebullah CM and Mishra SC (197! C 3 polymorphism in some Indian populations. Hum Hered *29* : 193-196.

49. Pingle U (1975) A comparative study of mating systems and marriage distanc patterns between five tribal groups of Utnur Taluka, Adilabad District of Andhr Pradesh. Proc 2nd Ann Conf Indian Soc Hum Genet, Calcutta, pp 1-15.

50. Ramesh A, Blake NM, Vijayakumar M and Murty JS (1980) Genetic studies o the Chenchu tribe of Andhra Pradesh, India. Hum Hered *30* : 291-298.

51. Ramesh A, Murty JS and Blake NM (1979) Genetic studies on the Kolams o Andhra Pradesh, India. Hum Hered *29* : 147-153.

52. Rao PM, Blake NM and Veerraju P (1978) Genetic studies on the Savara an Jatapu Tribes of Andhra Pradesh. Hum Hered *28* : 122-131.

53. Rao PR and Goud JD (1979) Sickle-cell haemoglobin and glucose-6-phosphat dehydrogenase deficiency in tribal populations of Andhra Pradesh. Indian J Mec Res *70* : 807-813.

54. Rao PR, Goud JD and Swamy BR (1979) Serum albumin variants from populations of Andhra Pradesh, S. India. Hum Genet *51* : 221-224.

55. Roberts DF, Papiha SS, Rao GN, Habeebullah CM, Kumar N and Murty KJR (1980) A genetic study of some Andhra Pradesh populations. Ann Hum Biol *7* : 199-212.

56. Rothman ED and Adams J (1978) Estimation of expected number of rare alleles of a locus and calculation of mutation rate. Proc Natl Acad Sci USA *75* : 5094-5098.

57. Saha N, Kirk RL, Shanbhag S, Joshi SR and Bhatia HM (1974) Genetic studies among the Kadar of Kerala. Hum Hered *24* : 198-218.

58. Saha N, Kirk RL, Shanbhag S, Joshi SR and Bhatia HM (1976) Population genetic studies in Kerala and the Nilgiris (South West India). Hum Hered *26* : 175-197.

59. Santachiara-Benerecetti SA, Cattaneo A and Meera Khan P (1972) A new variant allele AK^5 of the red cell adenylatekinase polymorphism in a non-tribal Indian population. Hum Hered *22* : 171-173.

60. Santchiara-Benerecetti SA, Cattaneo A and Meera Khan P (1972) Rare phenotypes of PGM_1 and PGM_2 loci and a new PGM_2 variant allele in the Indians. Amer J Hum Genet *24* : 680-685.

61. Shaw CR (1965) Electrophoretic variation in enzymes. Science *149* : 936-943.

62. Siciliano MJ, Bordelon MR and Kohler PO (1978) Expression of human adenosine deaminase after fusion of adenosine deaminase deficient cells with mouse fibroblasts. Proc Natl Acad Sci USA *75* : 936-940.

63 Takahata N (1980) Composite stepwise mutation model under the neutral mutation hypothesis. J Mol Evol *15* : 13-20.

64. Tchen P, Seger J, Bois E, Grenand F, Fribourg-Blanc A and Feingold N (1978) A genetic study of two French Guiana Amerindian populations II : rare electrophoretic variants. Hum Genet *45* : 317-326.

65. Veerraju P. Consanguinity in tribal communities of Andhra Pradesh. In : Verma IC,

ed. *Medical Genetics in India*, Vol. 2. Pondicherry : Auroma, 1978; 157-164.

66. Voelker RA, Langley CM, Leigh-Brown AJ, Ohnishi S, Dickson B, Montgomery E and Smith SE (1980) Enzyme null alleles in natural populations of *Drosophila melanogaster* : frequencies in a North Carolina population. Proc Natl Acad Sci USA *77* : 1091-1095.
67. Voelker RA, Schaffer HE and Mukai T (1980) Spontaneous allozyme mutations in *Drosophila melanogaster* : rate of occurrence and nature of the mutants. Genetics *94* : 961-968.
68. Wright S (1931) Evolution in Mendelian populations. Genetics *16* : 97-159.
69. Wright S (1946) Isolation by distance in diverse systems of mating. Genetics *31* : 39-59.

Dr. K. K. Bhatia
Department of Human Biology
John Curtin School of Medical Research
P. O. Box 334
Canberra, A. C. T., 2601
Australia

Am J Hum Genet 33:112–122, 1981

Frequency of Private Electrophoretic Variants and Indirect Estimates of Mutation Rate in Papua New Guinea

K. K. Bhatia,[1] N. M. Blake, S. W. Serjeantson, and R. L. Kirk

SUMMARY

Data on rare and private electrophoretic variants have been used to estimate mutation rates for populations belonging to 55 language groups in Papua New Guinea. Three different methods yield values of 1.42×10^{-6}, 1.40×10^{-6}, and 5.58×10^{-6}/locus per generation. The estimates for three island populations off the north coast of New Guinea—Manus, Karkar, and Siassi—are much lower. The variability in mutation rates estimated from rare electrophoretic variants as a function of population size is discussed. The mean mutation rate in Papua New Guinea is less than half the estimates obtained for Australian Aborigines and Amerindians.

INTRODUCTION

In a previous paper [1], we gave data for the frequency in Australian Aborigines of private electrophoretic variants for enzymes controlled by 25 loci, and these data were used to determine an indirect estimate of the mutation rate. Three different methods yielded values of 6.11×10^{-6}, 2.78×10^{-6}, and 12.86×10^{-6}/locus per generation for the total sample of Aborigines, and similar values were obtained for a series drawn from one tribal population in Australia.

Neel and Neel and Rothman initiated such studies using data for South American Indian populations [2, 3], and the same data were re-examined by Nei [4]. The mutation rate estimates for these populations are comparable to those obtained by us in Australian Aborigines. Tchen et al. [5] estimated mutation rates for Amerindian populations in French Guiana, and Chakraborty and Roychoudhury [6] have done the same for some South Asian populations.

Received December 31, 1979; revised May 28, 1980.

[1] Department of Human Biology, John Curtin School of Medical Research, Canberra, A.C.T. 2601, Australia.

© 1981 by the American Society of Human Genetics. 0002-9297/81/3301-0013$02.00

We have now extended our own studies to include data for 21 protein loci over 55 speech communities in Papua New Guinea. It is hoped that these new data will assist in understanding some of the complex factors that influence indirect estimates of the mutation rate in man.

STUDY POPULATION

Papua New Guinea comprises the portion of the island of New Guinea east of longitude 141° east plus several geographically related islands, including New Britain, the Admiralty Islands, and Bougainville. The census size of Papua New Guinea is approximately three million, or about 67% of the estimated total Melanesian population, and its population is one of the most complex linguistically and most socially fragmented areas of the world. It is estimated that there are about 700 speech communities in Papua New Guinea, divided between two major linguistic language types: Papuan and Austronesian [7]. A survey of the patterns of social structure is given in the *Encyclopaedia of Papua and New Guinea* [8].

The present analysis is based on samples collected from populations belonging to 47 languages, also called speech communities, on the mainland of Papua New Guinea, together with two speech communities from Karkar Island and five from Siassi Islands (both off the northern shore of New Guinea), and one speech community, Titan, from the Great Admiralty Island, also called Manus. The populations of these three offshore islands (Karkar, Manus, and Siassi), although having evolved in a similar ecological setting, have been exposed to different types of population pressures. These societies, like the coastal regions of mainland Papua New Guinea, have been at the crossroads of migrations in and around the Pacific, and may well have had their genetic composition considerably altered through repeated contact with outsiders. The populations of these three islands have also been analyzed separately for respective mutation rates.

LABORATORY DATA

The samples analyzed were collected during the past 12 years by us or by collaborators involved in medical surveys in Papua New Guinea. Material in all cases was shipped by air to Canberra, and testing was carried out in our laboratories using standard procedures outlined in Blake et al. [9]. The variants of a few other systems were tested using the techniques as follows: GPT [10], ESD [11], CA_1 and CA_2 [12], GLO [13], and PGM_1 and PGM_2 [14]. The list of enzymes studied, along with their sample size and subunit molecular weights, are given in table 1.

Seven of the 21 loci in the study are polymorphic, and six of the 21 are invariant. Out of the 53 alleles segregating, 24 alleles are rare as a whole. Two other alleles, namely, PGM_1^3 and PGM_2^9, although polymorphic, are considered here to be of New Guinean origin. This raises the number of variants to 26. The names of these variants, along with their frequencies, are given in table 2.

Four of these 26 variant alleles (*Hb* $J^{\text{Tongariki}}$, GOT^3, GPT^3, and GPT^6) cannot be assigned with certainty to Papua New Guinea because of their presence in appreciable numbers in other Western Pacific populations that, as is true for the Japanese during World War II, have had contact with Papua New Guinea in the past. The

TABLE 1

GENETIC MARKERS IN PAPUA NEW GUINEANS

LOCUS NO.	ENZYME SYSTEM	ABBREVIATION	SUBUNIT SIZE (IN DALTONS)	SAMPLE SIZE			
				Total population	Karkar Island	Manus Island	Siassi Islands
	Multimers:						
1	Hemoglobin-α	Hbα	15,000	6,874	1,086	80	287
2	Hemoglobin-β	Hbβ	16,000	6,874	1,086	80	287
3	Superoxide dismutase	SOD_A	16,000	7,322	1,091	183	287
4	Glyoxalase I	GLO	24,000	2,192	...	...	283
5	Esterase D	ESD	28,000	5,453	870	...	286
6	Malate dehydrogenase-2	MDH_2	35,000	7,856	1,091	183	286
7	Lactate dehydrogenase A	LDH_A	35,000	7,858	1,091	183	286
8	Lactate dehydrogenase B	LDH_B	35,000	7,858	1,091	183	286
9	Glutamic-oxaloacetic transaminase	GOT	46,000	4,441	433	...	...
10	Isocitrate dehydrogenase	ICD_S	48,000	4,908	433	...	287
11	Glutamic-pyruvic transaminase	GPT	50,000	5,205	837	...	93
12	6-Phosphogluconate dehydrogenase	6PGD	53,000	7,809	1,091	183	284
13	Phosphohexose isomerase	PHI	62,000	6,934	617	183	221
	Monomers:						
14	Acid phosphatase-1	ACP_1	15,000	7,421	1,090	183	287
15	Adenylate kinase-1	AK_1	22,000	5,551	617	183	...
16	Carbonic anhydrase-1	CA_1	29,000	1,857	...	...	287
17	Carbonic anhydrase-2	CA_2	29,000	1,370	...	...	...
18	Phosphoglycerate kinase	PGK	50,000	7,818	1,090	183	260
19	Phosphoglucomutase 1	PGM_1	51,000	7,775	1,087	183	283
20	Peptidase B	PEPB	55,000	5,108	617	183	...
21	Phosphoglucomutase 2	PGM_2	61,000	7,787	1,086	183	283
Total				126,271	16,404	2,356	4,573

TABLE 2

NO. AND FREQUENCIES OF PRIVATE VARIANTS IN PAPUA NEW GUINEA

SERIAL NO.	ENZYME	PRIVATE VARIANT	TOTAL POPULATION		KARKAR ISLAND		MANUS ISLAND		SIASSI ISLAND	
			NO. COPIES	% GENE FREQUENCY	NO. COPIES	% GENE FREQUENCY	NO. COPIES	% GENE FREQUENCY	NO. COPIES	% GENE FREQUENCY
1	Hbα*	*Hb J*$^{\text{Tongariki}}$	46	0.34	38	1.75	0	0.00	7	1.22
2	ESD	*ESD*^{3}	1	0.01	0	0.00	...	...	0	0.00
3	MDH_2	*MDH*$_2$^{3}	106	0.68	16	0.73	0	0.00	0	0.00
4	MDH_2	*MDH*$_2$^{6}	1	0.01	0	0.00	0	0.00	0	0.00
5	LDH_A	*LDH*$_A$$^{\text{Wantoat}}$	1	0.01	0	0.00	0	0.00	0	0.00
6	LDH_A	*LDH*$_A$$^{\text{WH}}$	9	0.06	0	0.00	0	0.00	0	0.00
7	LDH_B	*LDH*$_B$$^{\text{KK2}}$	1	0.01	1	0.05	0	0.00	0	0.00
8	LDH_B	*LDH*$_B$$^{\text{KK3}}$	1	0.01	1	0.05	0	0.00	0	0.00
9	GOT*	*GOT*^{3}	1	0.01	0	0.00	...	...	...	...
10	GPT*	*GPT*^{3}	8	0.08	0	0.00	...	...	0	0.00
11	GPT*	*GPT*^{6}	10	0.10	3	0.18	...	...	0	0.00
12	6PGD	*PGD*$^{\text{Wantoat}}$	38	0.24	3	0.14	0	0.00	0	0.00
13	6PGD	*PGD*$^{\text{Hackney}}$	1	0.01	0	0.00	0	0.00	0	0.00
14	PHI	*PHI*^{2}	1	0.01	0	0.00	0	0.00	1	0.23
15	PHI	*PHI*^{3}	1	0.01	0	0.00	0	0.00	0	0.00
16	PHI	*PHI*^{5}	4	0.03	0	0.00	0	0.00	0	0.00
17	PHI	*PHI*^{9}	2	0.02	0	0.00	0	0.00	0	0.00
18	PHI	*PHI*^{11}	1	0.01	0	0.00	0	0.00	0	0.00
19	PGK	*PGK*^{2}	18	0.12	0	0.00	0	0.00	11	2.12
20	PGK	*PGK*^{4}	62	0.40	5	0.23	0	0.00	3	0.58
21	PGM_1	PGM_1^{3}	254	1.63	17	0.78	0	0.00	1	0.18
22	PGM_1	*PGM*$_1$^{6}	3	0.02	0	0.00	0	0.00	0	0.00
23	PEPB	*PEPB*^{2}	1	0.01	0	0.00	0	0.00	0	0.00
24	PGM_2	*PGM*$_2$^{3}	1	0.01	0	0.00	0	0.00	1	0.18
25	PGM_2	*PGM*$_2$^{9}	903	5.80	166	7.64	0	0.00	10	1.77
26	PGM_2	*PGM*$_2$^{10}	112	0.72	4	0.18	0	0.00	3	0.53

* Widely distributed in Western Pacific region.

introduction of these variants to Papua New Guinea through admixture, therefore, cannot be excluded. Thus, we have 22 alleles at 21 loci that can be regarded as indigenous to Papua New Guinea.

On Karkar Island we detected 10 of the 26 allelic variants listed in table 2 over a set of 18 enzyme loci. Only two of these, namely, LDH_B^{KK2} and LDH_B^{KK3}, are unique to Karkar and are represented by single copies only. Seven of the other eight variants are mainland markers also; the only exception is $Hb\ J^{Tongariki}$, polymorphic on Karkar but rare on the mainland. This last allele also has wide distribution in other parts of Melanesia.

We did not find any private variant at 14 red cell enzyme loci tested on Manus. The absence of Southwest Pacific genetic markers, such as PGM_1^3, PGM_1^7, PGM_2^9, PGM_2^{10}, PGK^2, PGK^4, and $Hb\ J^{Tongariki}$, makes this population unique in the Western Pacific area.

For a set of 17 loci, we came across two unique variants on Siassi Islands, namely, PHI^2 and PGM_2^3, with a single copy each. The population of Siassi Islands, however, has a fair proportion of markers distributed in the Western Pacific region. Except for a very low frequency of PGM_1^3, four other variants, namely, PGK^2, PGK^4, PGM_2^{10}, and $Hb\ J^{Tongariki}$, are in polymorphic proportions on these islands.

ESTIMATION OF I, I_q, AND $\hat{I}$

Three different estimates for the mean number of variants/locus were calculated as given by Kimura and Ohta [15], Nei [4], and Rothman and Adams [16]. For the 21 loci in the present study, we have detected 22 unique (20 rare and two polymorphic) variants that give values of I, I_q, and $\hat{I}$ as 1.05, 0.95, and 1.78, respectively. The two parameters of geometric distribution involved in the estimation of $\hat{I}$, namely, b and c ($\hat{b} = 0.5567$; $\hat{c} = 0.3865$) [16], have been estimated from the data for Kiunga in the Western Province given by Serjeantson [17].

Considering the islands separately, the estimates of I, I_q, and $\hat{I}$ are 0.11, 0.11, and 0.46, respectively, for Karkar Island. The respective values for Siassi Islands are 0.12, 0.12, and 0.86, and for Manus Island, zero for each estimate. The parameters of geometric distribution, b and c, used in the estimation of transition probabilities for both Karkar and Siassi populations ($\hat{b} = 0.2711$; $\hat{c} = 0.7126$) were calculated from the data on Karkar Island by Hornabrook [18].

Since the calculation of I depends a priori on the observed distribution of copies of rare alleles, $\tilde{g}(i)$, and the ratio (f) of sample (n) to effective population size (N_e), these estimates are inflated by 4.09 and 8.57 times on Karkar and Siassi, respectively, when compared with the observed value of I. This increase, when compared with a less than twofold increase for a similar estimate for the total sample, is highly inflated. It would seem that estimations of $\hat{I}$ by Rothman and Adams's [16] method gives reliable results only with large absolute sample sizes.

ESTIMATION OF N_e

In the absence of historical records, it is difficult to estimate accurately the average effective population size (N_e) of linguistic groups in Papua New Guinea. However, since the estimates of mutation rate are highly dependent on the estimate

of N_e, we discuss this in some detail. The Papua New Guinea Bureau of Statistics census of 1971 reported a total population of 2,435,409 indigenous persons, of whom 41.6% were in the reproductive age group of 15–44 years, with a similar proportion (41.5%) married at least once. With a minimum of 700 documented language groups [7], the maximum estimate of language group effective size is 1,447.

This maximum value of N_e may be considered a gross overestimate of population size during past generations. Van de Kaa [19] considers that the Papua New Guinean population was stable between 1890 and 1939, partly because there is no evidence to suggest otherwise, but mainly because analysis of the few surveys undertaken at that time show little demographic change. In our calculations, we assume that actual population size during the last 5–10 generations more closely approximates the census figures of 1939 and that the population of 1939 was very close to 50% of that enumerated in 1971.

In Papua New Guinea, estimates of N_e of language groups also require correction for the extreme variability in language group size. Linguistic groups may comprise fewer than 100 persons, as in Gorovu in the Ramu phylum [20], or more than 150,000 persons, as in Enga in the Western Highlands [21]. By far, the majority of language groups have fewer than 5,000 speakers. Since the average value of N_e more closely approximates the harmonic than the simple mean of language group size, we have analyzed the three main linguistic phyla represented in the Madang Province to estimate the ratios of the harmonic means ($\overline{H}$) to simple means ($\overline{N}$). For the Adelbert Range phylum, $\overline{H}/\overline{N}$ is 35%; for the Ramu phylum, 33%; and for the Madang phylum, 40%. The combined value for 80 languages is 36%.

Therefore, for estimation of the mean number of speakers per language, we take 50% of 2,435,409 as the total population prior to 1939, distributed among 700 languages of varying sizes, with an average of 1,740 speakers. Since the harmonic mean of language group size is 36% of the simple mean, the more appropriate estimate is 626 speakers per language when correction is made for variability in language group size.

The effective population size is further modified by the proportion in the reproductive age group, variability in fertility, and deviation of the sex ratio from 1:1. The adult sex ratio was less than unity in the 1971 census and greater than unity in the previous census of 1966 [22]; so we shall assume the sex ratio in the reproductive age group fluctuates around 1:1. The proportion in the reproductive age group is more difficult to estimate accurately. In 1971, 41.6% of the population was aged between 15 and 44 years, compared with 45.0% in 1966 [22], and Serjeantson [23] recorded that 49% of the population of two relatively unacculturated Papua New Guinean language groups was aged between 16 and 45 years. We believe that the proportion in the reproductive age group in past generations was closer to 49%, the estimate we used in our calculations, than to the 42% currently observed.

Variation in fertility will modify N_e if the index of variability ($V/\bar{k}$) deviates from unity [24], $\bar{k}$ and V being the mean number and variance of surviving offspring, respectively. In Papua New Guinea, the index of variability is inflated by factors such as polygyny, which was reported by 9% of married males as recently as 1971 [22]. Serjeantson [23] estimated the index of variability as 1.22 in males from the

Yonggom group with 10% polygyny, and 2.09 in males from an additional group (Awin) with 28% polygyny. The corresponding values in females were 0.96 and 1.40 in a population with such comparatively low fertility [17] that it may well reflect the demographic structure of most Papua New Guinean groups prior to 1939.

With an average index of variability of 1.4 and 49% of the population in the reproductive age group, the ratio of $N_e v/N_e$ is 83.7%. The average effective size of language groups in Papua New Guinea is estimated as 49% of 626, or 307 persons, and this is the value used in estimating the mean survival time for fresh mutations in Papua New Guinean language groups. In general, it is the language groups with a relatively large number of speakers that have been sampled, so that the average effective size of language groups with genetic data available exceeds slightly the average size of language groups in Papua New Guinea as a whole. Making similar adjustments as above for rapid population expansion in the last generation, for variation in language sizes, and for the proportion in the reproductive age group, the total effective population size for the 55 languages in this series is 34,450. We use this value in all our calculations.

The sizes of the three island populations (Karkar, Manus, and Siassi) stood at 9,110, 13,839, and 4,715, respectively, in 1937–1939 [19], with 50.3%, 62.7%, and 59.5% in the adult age group. After adjusting for the proportion in the reproductive age group, polygyny and sex ratio values of N_e are 3,735, 6,805, and 2,310 for Karkar, Manus, and Siassi, respectively.

ESTIMATION OF $\bar{t}_o$

The mean survival time for fresh mutations that will ultimately be lost from the population ($\bar{t}_o$) was given by Kimura and Ohta [15] and Nei [25]. This value is estimated for a Papua New Guinean language group as: $\bar{t}_o = 2\,N_e v/N_e \ln(2N_e) = 2 \times (0.837) \ln(2 \times 307) = 10.74$ generations, which is different from the estimate given by Li and Neel [26] of 5.71 from simulation studies of Amerindian populations. The estimates for Karkar and Siassi were calculated to be 14.92 and 14.12 generations, respectively. We have used these estimates for generating mutation rates by Kimura and Ohta's method.

ESTIMATION OF MUTATION RATES

The estimation of mutation rates has been carried out using three indirect methods as mentioned above. Table 3 shows these estimates for the total Papua New Guinean population. The three estimates of μ by the methods of Kimura and Ohta [15], Nei [4], and Rothman and Adams [16] are 1.42×10^{-6}, 1.40×10^{-6}, and 5.58×10^{-6}/locus per generation, respectively. These estimates range from approximately 23% to 50% of similar estimates obtained for the Australian Aborigines by Bhatia et al. [1].

Neel and Rothman [3] have used the estimate for $\hat{I}$ for the average number of mutant variants per locus in the formulation of Kimura and Ohta [15], instead of the observed value of I. A similar adjustment for differences between sample size (n) and effective population size (N_e) may be made in Nei's formulation, as:

$$\mu = \frac{\hat{I}q}{4N_e \ln (2N_e q)},$$

given that

$$\hat{I}_q = \frac{I_q}{1 - \Sigma g(j)(1-f)^j},$$

where $q = .01$, $f = n/N_e$, j is the number of copies, and $\tilde{g}(j)$ is the observed proportion of variants with j copies in the frequency distribution of rare variants (frequency less than .01) only. These modifications yield the new estimates as $\mu = 2.36 \times 10^{-6}$ and 2.38×10^{-6}/locus per generation for Kimura and Ohta's [15] and Nei's [4] methods, respectively.

The estimates of mutation rate for island populations show a wide range. The value of μ on Manus for a set of 14 protein loci is zero. The estimates of μ for Karkar and Siassi are given in table 3. Estimating the total number of variants in the populations with limited observations is highly unreliable, as is seen from the results for mutation rates in these populations by Rothman and Adams's [16] method. The estimates of μ obtained by the methods of Kimura and Ohta [15] and Nei [4] on these islands are, however, comparable to similar estimates generated for the Waljbiri tribe in Australian Aborigines [1].

DISCUSSION

The estimates of mutation rates as obtained from a set of protein loci are affected seriously by a number of factors. Probably the most controversial aspect of these indirect estimates is the estimation of effective population size (N_e). This is particularly difficult in the Papua New Guinean communities that have recently been undergoing tremendous demographic changes. The impact of recent population expansion can be judged from the high proportion of private polymorphisms with limited geographical distributions. Out of 26 variants detected, as many as 10 have attained polymorphic proportions in various Papua New Guinean communities, six of them in the highlands, one on both Karkar and Siassi, and three in both highland and coastal communities.

The role of sample size and subunit size in affecting the detection and introduction of rare variants has been stressed by a number of authors; for example, Nei et al. [27] and Bhatia [28]. In the present study, the mean sample sizes for loci with and

TABLE 3

MUTATION RATES IN PAPUA NEW GUINEA

POPULATION	$\mu \times 10^6$		
	Kimura and Ohta's method	Nei's method	Rothman and Adams's method
Total.....	1.42	1.40	5.58
Karkar Island.....	0.98	2.57	26.40
Manus Island	0.00	0.00	0.00
Siassi Islands.....	1.84	7.58	94.56

without variants are 7,226 and 4,955, respectively. This difference emphasizes the need for a sample size of at least 3,000, as suggested by Eanes and Koehn [29], before any attempts are made to generate mutation rates. Similarly, the mean subunit size for loci with these variants is 46,300 daltons compared with 28,180 for the invariant loci. It is thus important to make comparisons of mutation rates only among populations with similar mean sample sizes and mean subunit sizes.

The effect of sample size on the mean number of rare variants per locus may be seen in a comparison of the data on Papua New Guinean communities with the data on the Australian Aborigines. While there is similarity with respect to protein loci included in the two studies (mean molecular weights of the subunits are 36,869 and 37,560 daltons for Papua New Guineans and Australian Aborigines, respectively), the differences in the mean per locus sample sizes of 6,036 in the Papua New Guineans and 2,607 in the Australian Aborigines are reflected in the respective estimates of 1.05 and 0.64 for I.

However, the mean number of rare variants/locus per individual (I/n) is higher in Australian Aborigines (2.46×10^{-4}) in comparison with Papua New Guinean communities (1.74×10^{-4}). Since the mean number of electromorphs recovered is a logarithmic function of sample size and the distribution of electromorphs is skewed further with sample size increase, it will be appropriate to compute the mean number of rare variants/locus per individual only in terms of effective population size (N_e), rather than in terms of sample size (n). The two estimates for Australian Aborigines and Papua New Guineans then become 6.99×10^{-5} and 3.05×10^{-5}, respectively, a difference of 2.29 times.

Nei and Chakraborty [30] have shown that the mean number of silent alleles, undetectable by electrophoresis, that contribute to an electromorph is higher in populations with large $N_e\mu$'s than in populations in which this is small. On the basis of this argument, the proportion of mean numbers of silent alleles is likely to be much higher in Papua New Guinean communities than in Australian Aborigines. Since the mean number of rare variants/locus (I) reflects the incidence of mutation rate in a population, the ratio of N_eI in these two populations, when adjusted for sample sizes, yields a value of 2.66 times more silent alleles in Papua New Guineans than Australian Aborigines. The results at electromorph level (notwithstanding the differences in mutation rate at codon level between the two populations) are almost negligible.

Because of these various factors that may affect the mean number of rare variants per locus, the indirect estimates of mutation rate will show similar variations. It is not surprising, therefore, that estimates of μ generated from protein data for Papua New Guinea differ about twofold from estimates generated on a similar scale for Amerindians by Neel and Rothman [3] and for Australian Aborigines by Bhatia et al. [1]. The estimate of μ for a group of tribes in India, however, is lower by more than an order of magnitude compared with these estimates. This may be because of a recent population increase in India and differences in sample size, in the number of loci studied, and in technical methods employed in blood collection, none of which have been taken into consideration by Chakraborty and Roychoudhury [6].

The estimates of Nei's μ for individual populations in South America (Neel and Rothman [3] and Tchen et al. [5]), in India (Chakraborty and Roychoudhury [6]), in Australia (Bhatia et al. [1]), and in Papua New Guinea (present study), however, show a wide variation with a mean estimate of 1.19×10^{-5}/locus per generation. The differences range from $0 - 9.28 \times 10^{-5}$, with a more or less equal number of populations with estimates of the order of 10^{-7}, 10^{-6}, and 10^{-5}. Our results for Karkar, Manus, and Siassi are at the lower end of this distribution, which conforms with the lower estimate found for Papua New Guinean populations treated as a whole.

REFERENCES

1. Bhatia K, Blake NM, Kirk RL: The frequency of private electrophoretic variants in Australian Aborigines and indirect estimates of mutation rate. *Am J Hum Genet* 31:731–740, 1979
2. Neel JV: Private genetic variants and the frequency of mutation among South American Indians. *Proc Natl Acad Sci USA* 70:3311–3315, 1973
3. Neel JV, Rothman ED: Indirect estimates of mutation rates in tribal Amerindians. *Proc Natl Acad Sci USA* 75:5585–5588, 1978
4. Nei M: Estimation of mutation rates from rare protein variants. *Am J Hum Genet* 29:225–232, 1977
5. Tchen P, Séger J, Bois E, Grenand F, Fribourg-Blanc A, Feingold N: A genetic study of two French Guiana Amerindian populations II: Rare electrophoretic variants. *Hum Genet* 45:317–326, 1978
6. Chakraborty R, Roychoudhury AK: Mutation rates from rare variants of proteins in Indian tribes. *Hum Genet* 43:179–183, 1978
7. Wurm SA: Language distribution in the New Guinea area, in *New Guinea Area Languages and Language Study*, vol I, Pacific Linguistics, edited by Wurm SA, series C, no. 38, Canberra, Australian National Univ., 1975, pp 3–38
8. Lepervenche M: Social structure, in *Encyclopaedia of Papua and New Guinea*, edited by Ryan D, Australia, Melbourne Univ. Press, 1972
9. Blake NM, Kirk RL, McDermid EM, Case J, Bashir H: The distribution of blood, serum protein and enzyme groups in a series of Lebanese in Australia. *Aust J Exp Biol Med Sci* 51:209–220, 1973
10. Chen S-H, Giblett ER, Anderson JE, Fossum BLG: Genetics of glutamic pyruvic transaminase: its inheritance, common and rare variants, population distribution and differences in catalytic activity. *Ann Hum Genet* 35:401–409, 1972
11. Hopkinson DA, Mestriner MA, Cortner J, Harris H: Esterase D: a new human polymorphism. *Ann Hum Genet* 37:119–137, 1973
12. Hopkinson DA, Coppock JS, Muhlemann MF, Edwards YH: The detection and differentiation of the products of the human carbonic anhydrase loci, CAI and CAII, using flurogenic substrates. *Ann Hum Genet* 38:155–162, 1974
13. Kömpf J, Bissbort S, Gussmann S, Ritter H: Polymorphism of red cell glyoxalase I (E.C.: 4.4.1.5): a new genetic marker in man. *Humangenetik* 27:141–143, 1975
14. Blake NM, Omoto K: Phosphoglucomutase types in the Asian-Pacific area: a critical review including new phenotypes. *Ann Hum Genet* 38:251–273, 1975
15. Kimura M, Ohta T: The average number of generations until extinction of an individual mutant gene in a finite population. *Genetics* 63:701–709, 1969
16. Rothman ED, Adams J: Estimation of expected number of rare alleles of a locus and calculation of mutation rates. *Proc Natl Acad Sci USA* 75:5094–5098, 1978
17. Serjeantson S: Marriage patterns and fertility in three New Guinean populations. *Hum Biol* 47:399–413, 1975

18. HORNABROOK RW: The demography of the population of Karkar Island. *Philos Trans R Soc Lond [Biol]* 268(B):229–239, 1974
19. VAN DE KAA DJ: *The Demography of Papua New Guinea's Indigenous Population*. Ph.D. thesis, Canberra, Australian National Univ., 1971–72
20. Z'GRAGGEN JA: *Classificatory and Typological Studies in Languages of the Madang District*. Pacific Linguistics, Series C, no. 19, Canberra, Australian National Univ., 1971, pp 1–179
21. WURM SA: Eastern central trans-New Guinea phylum languages, in *New Guinea Area Languages and Language Study*, vol I, Pacific Linguistics, edited by WURM SA, series C, no. 38, Canberra, Australian National Univ., 1975, pp 461–524
22. TERRITORY OF PAPUA NEW GUINEA BUREAU OF STATISTICS: *Papua New Guinea, Summary of Statistics*. Port Moresby, Government Printer, 1971–72
23. SERJEANTSON S: *The Population Genetic Structure of the North Fly River Region of Papua New Guinea*. Ph.D. thesis, Univ. of Hawaii, 1970
24. CROW JF, MORTON NE: Measurement of gene frequency drift in small populations. *Evolution* 9:202–214, 1955
25. NEI M: Extinction time of deleterious mutant genes in large populations. *Theor Popul Biol* 2:419–425, 1971
26. LI FHF, NEEL JV: A simulation of the fate of a mutant gene of neutral selective value in a primitive population, in *Computer Simulation in Human Population Studies*, edited by DYKE B, MACCLUER JW, New York, Academic Press, pp 221–240, 1974
27. NEI M, CHAKRABORTY R, FUERST PA: Infinite allele model with varying mutation rate. *Proc Natl Acad Sci USA* 73:4164–4168, 1976
28. BHATIA K: Factors affecting electromorph mutation rates in man: an analysis of data from Australian Aborigines. *Ann Hum Biol* 7:45–54, 1980
29. EANES WF, KOEHN RK: Relationship between subunit size and number of rare electrophoretic alleles in human enzymes. *Biochem Genet* 16:971–985, 1978
30. NEI M, CHAKRABORTY R: Electrophoretically silent alleles in a finite population. *J Mol Evol* 8:381–385, 1976

Acta Anthropogenetica, Vol. 5 (1) : 41-56 (1981)

Genetic Studies on Some Tribes of the Telangana Region, Andhra Pradesh, India

N.M. BLAKE[1], A. RAMESH[2]*, M. VIJAYAKUMAR[2], J.S. MURTY[2]** AND K. K. BHATIA[1]

1. Department of Human Biology, John Curtin School of Medical Research, Canberra.
2. Department of Genetics, Osmania University, Hyderabad.

Key words

Tribes. Andhra Pradesh. Blood groups. Enzyme groups. Genetic distance.

Abstract

Phenotype distributions and gene frequencies of nine red cell enzyme systems and haemoglobin are presented for six tribal populations from the Telangana region of Andhra Pradesh. ABO, MN and Rh blood group data are presented for four of these tribes. The results have been compared with those from other Andhra Pradesh tribal Populations. The Yerukula tribe are notable for the presence of PGM_1^7 at polymorphic frequency, the occurrence of a single example of $PGM_2^{1^0}$ and the absence of Hb^S.

Genetic distance comparisons were made for the six tribal populations reported in this paper and five others taken from the literature.

Introduction

The tribal populations of Andhra Pradesh are comprised of some 32 Aboriginal groups which vary in size, geographical distribution, antiquity and cultural diversity. They constitute about 4% of the total population of well over 40 millions, and the object of this paper is to present data on a number of red cell enzyme and blood group systems for six tribal groups from the upland Telangana region of Andhra Pradesh: the Rajgonds and Pardhans from Adilabad District, the Koyas and Konda Reddis from

* Present address : Department of Genetics, Institute of Basic Medical Sciences, University of Madras, Madras-600 042, India.

** Present address : Department de Mathematiques, Institut des Sciences Exactes, Université de Constantine, Constantine, Algerie.

Khamman District and the Yerukula and Sugali from Mahabubnagar and Kurnool Districts.

The Gonds are the predominant tribal group in Adilabad District, to where they are reputed to have migrated from the neighbouring state of Maharashtra. Their language is Gondi, a Dravidian dialect, which is closer to Tamil and Kanarese than to Telegu. In the past, some families of Gonds occupied a high social stratum and were politically powerful rulers in the region known as Gondwana during the Moghul regime. These families adopted the name of Rajgonds and maintained a state of superiority over the other Gonds, but with their decline in political power the Rajgonds returned to equal social status with the Gonds.

The Pardhans are a minority group in Andhra Pradesh and are only found in Adilabad District. They were traditionally hereditary bards to the Gonds and their mother language is Marathi though they also use Gondi and Kolami.

The Koya are a large scheduled tribal population who inhabit forest as well as plains areas in Khamman District. They are settled agriculturalists like the Telegu peasants of the region and, in addition, they participate in some hunting and gathering in the forest areas, and in toddy tapping.

The Konda Reddi are a people whose physical make-up shows strong Veddoid influences. They are expert basket makers who subsist on cultivation as well as hunting and gathering.

Another tribal group who practice basket making are the Yerukula, whose language is a mixture of Tamil, Telegu and Kanarese. Whilst they are sometimes known by other names, such as Kaikadi in Telangana, their name Yerukula is derived from their profession of fortune-telling. The mythology of these people contains stories suggesting a relationship between the Yerukula and other tribes in Andhra such as the Chenchu and Yanadi.

The Sugalis, also called Lambadis, are a semi-nomadic tribe found in various states all over India, but particularly in the southern and western parts. They are of mixed origin and there are historical records which indicate their migration from the north, especially from Rajasthan and Gujarat, to southern India along with Moghul invaders for whom they acted as carriers of grain and other supplies.

In recent years some of the above six tribal groups have been studied for various genetic markers. Notable among such studies is that of Goud[8] which provided data on A_1A_2BO, Rh (D) blood group systems, ability to taste phenylthiocarbamide, colour blindness, Hp, Tf, Alb and Gm serum protein systems for the Rajgond, Pardhan, Naikpod and Koya

tribes of Adilabad, Karimnagar, Warangal and Khamman Districts. Similar data were also provided for the Yerukula tribe of Warangal and Hyderabad Districts by Goud and Rao[9]. Rao[20] presented data on the ABO blood group system in Konda Reddi, Koya and other tribal groups located in East Godavari District and Naidu and Veerraju[12] reported ABO and Rh (D) blood groups from tribal populations in Visakhapatnam

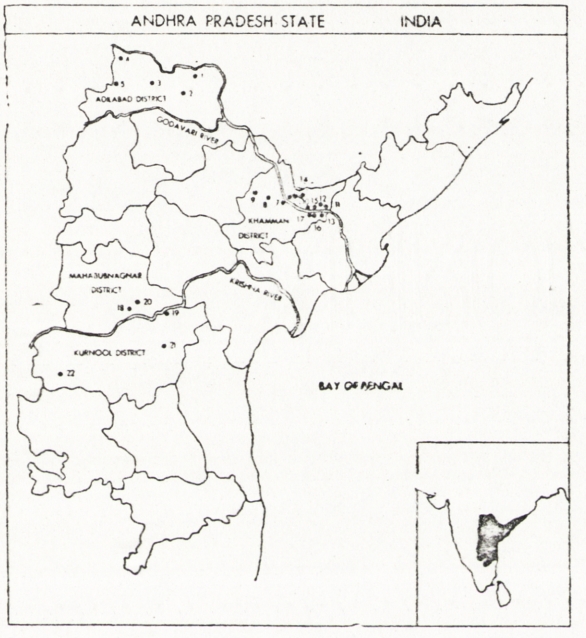

Fig. 1. Map of Andhra Pradesh showing the locations sampled, 1—Sirpur R (20), P (2). 2—Asifabad R (30), P (29). 3—Utnur R (47), P (17). 4—Adilabad R (9), P (42). 5—Boath R (30), P (9). 6—Badrachalam K (32). 7—Burgampad K (32). 8—Paloncha K (50). 9—Yellandu K (40). 10—V.R. Puram KR (3). 11—Kolluru KR (15). 12—Pocharam KR (26). 13—Koida KR (10). 14—Tekulloddi KR (13). 15—Jeedikuppa KR (13). 16—Katkuru KR (11). 17—Bheemavaram KR (1). 18—Lingala Y (26). 19—Srisailam Y (14). 20—Achempet S (7). 21—Mahanandi S (41) 22—Pattikonda S (52). R—Rajgond. P—Pardhan. K—Koya. KR—Konda Reddi. Y—Yerukula. S—Sugali. Figures in brackets refer to the number of blood samples collected.

District; ABO blood group studies on the Bagatha and Valmiki from the same District were carried out by Rao and Reddy[19]. Ramachandraiah[14] provided data on ABO blood group system, colour blindness and other anthropometric traits for the Lambadi (Sugali) tribe in Andhra Pradesh. Santachiara-Benerecetti *et al*[24, 25] studied some of the red cell enzyme systems on Konda Reddis, Koya Doras and some non tribals located in West Godavari District, whilst Goud and Rao[10] have provided data on serum protein systems for several tribal groups from Adilabad, Khamman and Warangal Districts. Two abnormal haemoglobins, Hb Rampa and Hb Koya Dora, have been reported in the Koya Dora[5, 6].

Data on a number of red cell enzyme systems for other tribal populations in Andhra Pradesh, have been provided by Rao *et al*[17] for the Savara and Jatapu, Ramesh *et al*[15] for the Kolams and Ramesh *et al*[16] for the Chenchu.

Materials and Methods

A total of 582 blood specimens were collected from school children in the Adilabad, Khamman, Mahabubnagar and Kurnool Districts of Andhra Pradesh and the locations sampled are shown in Fig. 1. For enzyme typing the blood was taken on to 3MM chromatography paper following the methods of Saha and Kirk[21] and Rao *et al*.[17] In some instances there was insufficient sample to test for all systems. For the blood grouping, samples were collected from fingerpricks into heparinised capillary tubes. No blood grouping was carried out for the Yerukula, and only the ABO and MN systems were studied for the Sugali. Where applicable the *Rh* gene frequencies were calculated using a maximum likelihood method.

Results and Discussion

The phenotype distribution for the ABO, MN and Rh systems are shown in Table I and the gene frequencies in Table II. The phenotype distributions and gene frequencies for haemoglobin and the red cell enzymes are listed in Table III. No variation was detected in malate dehydrogenase, lactate dehydrogenase or superoxide dismutase for any of the tribal groups reported in this paper.

Blood group systems

The results of the ABO and MN blood group systems for these six tribal populations show gene frequencies which are consistent with those for other tribal people from this region, though the value of 11.28% for A_1 in the Konda Reddi tends to be rather low and the value of 42.94% for

Table I. Distribution of blood group phenotypes for some Telangana tribal populations

ABO and MN systems		A_1	A_2	A_1B	A_2B	B	O		M	MN	N
Rajgonds	n=115	29	3	39	14	0	30	n=102	55	37	10
Pardhans	n=91	18	4	28	7	0	34	n=85	32	33	20
Koya	n=157	49	7	38	9	0	54	n=158	83	59	16
Konda Reddi	n=80	14	1	24	3	1	37	n=80	44	22	14
Sugali	n=100	30	0	37	8	0	25	n=100	55	38	7
Rh system		R_oR_o	R_1r	R_1R_1	R_2r	R_1R_2	R_1R_z	R_zR_z	$r'r'$	$r''r''$	rr
Rajgonds	n=107	5	18	72	1	10	0	1	0	0	0
Pardhans	n=67	1	16	39	2	7	1	0	0	0	1
Koya	n=135	1	35	92	0	3	0	0	1	1	1
Konda Reddi	n=74	0	13	59	0	2	0	0	0	0	0

Table II. Blood group allele frequencies for some Telangana tribal populations

Allele :	A_1	A_2	B	O	M	N	R_o	R_1	R_2	R_z	r	r'
Frequencies in %												
Rajgonds	20.77	1.32	26.52	51.39	72.06	27.94	13.72	80.20	4.97	1.11	0.00	0.00
Pardhans	14.76	2.23	21.52	61.48	57.06	42.94	4.85	75.97	6.57	0.90	11.72	0.00
Koya	20.67	2.27	16.41	60.65	71.20	28.80	5.80	77.22	1.12	0.00	9.12	6.74
Konda Reddi	11.28	1.26	19.37	68.09	68.75	31.25	0.00	89.86	1.35	0.00	8.78	0.00
Sugali	21.57	0.00	26.21	52.21	74.00	26.00	—	—	—	—	—	—

Table III. Distribution of phenotypes and gene frequencies for Hb and red cell enzyme groups for some Telangana tribal populations

		Rajgonds			Pardhans			Koya			Konda Reddi			Sugali			Yerukula		
		No.	Gene	%	No.	Gene	%	No.	Gene	%	No.	Gene	%	No.	Gene	%	No.	Gene	%
Hb		n=133			n=101			n=159			n=92			n=61			n=40		
	A	113	Hb^A	92 81	65	Hb^A	81.68	138	Hb^A	93.08	88	Hb^A	97.83	58	Hb^A	97.54	40	Hb^A	100.00
	AS	19	Hb^S	7 89	35	Hb^S	18.32	20	Hb^S	6.92	4	Hb^S	2.17	3	Hb^S	2.46	—		
	S	1			1			1			—			—			—		
PGM_1		n=134			n=88			n=159			n=92			n=60			n=39		
	1–1	66	PGM_1^1	69.40	42	PGM_1^1	67.61	76	PGM_1^1	66.67	51	PGM_1^1	71.20	37	PGM_1^1	78.33	17	PGM_1^1	66.67
	2–1	54	PGM_1^2	30 60	35	PGM_1^2	32.39	60	PGM_1^2	33.33	29	PGM_1^2	28.80	20	PGM_1^2	21.67	14	PGM_1^2	25.64
	2–2	14			11			23			12			3			3	PGM_1^7	7.69
	7–1	—			—			—			—			—			4		
	7–2	—			—			—			—			—			1		
PGM_2		n=134			n=88			n=159			n=92			n=60			n=40		
	1–1	134	PGM_2^1	100 00	88	PGM_2^1	100.00	159	PGM_2^1	100.00	92	PGM_2^1	100.00	60	PGM_2^1	100.00	39	PGM_2^1	98.75
	10–1	—			—			—			—			—			1	PGM_2^{10}	1.25
ACP_1		n=134			n=96			n=155			n=91			n=60			n=40		
	A	5	ACP_1^A	19.03	7	ACP_1^A	25.00	8	ACP_1^A	17.42	7	ACP_1^A	28.57	8	ACP_1^A	28.33	2	ACP_1^A	22.50
	AB	41	ACP_1^B	80.97	34	ACP_1^B	75.00	38	ACP_1^B	82.58	36	ACP_1^B	70.88	18	ACP_1^B	71.67	14	ACP_1^B	77.50
	B	88			55			109			46	APC_1^C	0.55	34			24		
	AC	—			—			—			1			—			—		
6-PGD		n=134			n=95			n=159			n=92			n=61			n=40		
	A	132	PGD^A	99.25	95	PGD^A	100.00	155	PGD^A	98.74	89	PGD^A	98.37	60	PGD^A	99.18	39	PGD^A	98.75
	AC	2	PGD^C	0.75	—			4	PGD^C	1.26	3	PGD^C	1.63	I	PGD^C	0.82	1	PGD^C	1.25
AK		n=134			n=101			n=159			n=92			n=61			n=40		
	1–1	115	AK^1	92.91	87	AK^1	92.08	138	AK^1	93.08	84	AK^1	95.65	47	AK^1	86.89	35	AK^1	93.75
	2–1	19	AK^2	7.09	12	AK^2	7.92	20	AK^2	6.92	8	AK^2	4.35	12	AK^2	13.11	5	AK^2	6.25
	2–2	—			2			1			—			2			—		
PHI		n=134			n=100			n=159			n=92			n=61			n=40		
	1–1	133	PHI^1	99.25	100	PHI^1	100.00	153	PHI^1	98.11	92	PHI^1	100.00	61	PHI^1	100.00	40	PHI^1	100.00
	3–1	1	PHI^3	0.75	—			6	PHI^3	1.89	—			—			—		

Table IV. Morton's θ_{ij} values (x10^4; upper triangular matrix) and Nei's D values (x10^4: lower triangular matrix) among 11 tribal populations from Andhra Pradesh

	Chenchu (M)	Chenchu (K)	Kolam	Savara	Jatapu	Koya	Rajgond	Pardhan	Konda Reddi	Lambadi	Yerukula
Chenchu (M)	X	149	159	131	135	186	177	300	150	125	214
Chenchu (K)	121	X	75	54	46	77	80	185	43	73	139
Kolam	90	62	X	63	46	28	36	89	49	69	152
Savara	88	72	27	X	25	48	74	196	49	35	131
Jatapu	87	73	18	9	X	41	72	198	31	65	128
Koya	129	57	19	32	31	X	35	138	52	68	161
Rajgond	129	56	43	75	76	24	X	111	63	68	171
Pardhan	139	60	47	84	81	68	67	X	175	187	309
Konda Reddi	123	45	55	62	53	41	54	68	X	62	138
Lambadi	50	101	48	54	66	74	84	111	94	X	138
Yerukula	55	56	30	28	30	46	65	89	59	31	X

N in the Pardhans tends to be high.

There is not a large amount of data available on the Rh system for populations in Andhra Pradesh with the exception of the Chenchu who have been studied extensively by Simmons *et al*[26] and Ramesh *et al*[16]. The frequencies of R_1 in the four populations reported in this paper are all rather higher than in the Chenchu with the R_2 values being correspondingly lower. The values for *r* in the Pardhans, Koya and Konda Reddis, are all similar to that in both groups of Chenchu noted above, though the allele was not detected in the Rajgonds. The Chenchu reported by Simmons *et al*[26] showed a high frequency of r; in this report, only in the Koya was this allele present, at a frequency of 6.7%.

Haemoglobin

Haemoglobin S is found commonly in the tribal groups of south India and Andhra Pradesh; Rao *et al*[17] have reported its presence in the Savara and Jatapu, Ramesh *et al*[15] for the Kolams, Rameh *et al*[16] for the Chenchu and Rao and Goud[18] for a number of other tribal groups. In the present study the Hb^S allele, confirmed by citrate-agar electrophoresis, was present in all groups except the Yerukula, the highest frequency being 18.3% in the Pardhans.

Red cell enzymes

Phosphoglucomutase, locus 1 :

Both the PGM_1^1 and PGM_1^2 alleles are present in all six tribal groups and the frequencies are within the range expected for southern Indian populations. However, of particular interest is the occurrence of the allele PGM_1^7 in the Yerukula where four examples of PGM_1 7-1 and one example of PGM_1 7-7 were observed to give a gene frequency of almost 8% for PGM_1^7, the highest value so far recorded anywhere in the world for this allele; however, the sample size is small and the two groups tested are probably inbred isolates. A frequency of almost 6% has been recorded for PGM_1^7 in Micronesians from the Western Carolines[1]. Santachiara-Benerecetti *et al*[24] have detected two examples of the phenotype PGM_1 7-1 in a non-tribal population from Andhra Pradesh and Ramesh *et al*[16] have reported a single PGM_1 7-1 in 139 Chenchu from Mahabubnagar.

Phosphoglucomutase, locus 2 :

The occurrence of variation at the PGM_2 locus has been reviewed by Blake and Omoto[2] and in that report the widespread distribution in New Guinea of the allele $PGM_2^{1^0}$ was described; also, the occurrence of a

single example of PGM_2 10-1 was noted in an Australian Aborigine from Arnhem Land. In India, Santachiara-Benerecetti *et al*[24] have reported PGM_2^1, $PGM_2^{6\,IND}$ and PGM_2^8 in non-tribals from Andhra Pradesh.

The occurrence of the phenotype PGM_2 10-1 in a single Yerukula individual is therefore of particular interest. The isozyme banding patterns, after electrophoresis in starch gels, for PGM_2 6^{IND}-1 and PGM_2 10-1 appear to be similar to each other, though the distribution of activity within the bands may be different as suggested by the Fig. 5 in Blake and Omoto[2]. However, whilst we have not been able to directly compare the New Guinea PGM_2 10-1 with PGM_2 6^{IND}-1, we have compared the Yerukula PGM_2 10-1 with the New Guinea variant, and believe them to be electrophoretically indistinguishable.

Acid phosphatase :

The frequency of ACP_1^A tends to be lower in south Indian populations than in those from the north; in the south, the range is from 3.5% for the Irulas[23] up to 38.1% in the Mahabubnagar Chenchu[16] and 46.7% in the Kota[7]. The frequencies for ACP_1^A in the six tribal groups of this paper are comparable with those for other tribal populations in Andhra Pradesh[15–17]. Of interest, however, is the occurrence of a single ACP_1 AC phenotype in a Konda Reddi. The ACP_1^C allele occurs, sporadically in Indian populations but seems to be even more uncommon in tribal populations. In south India, the only other report of $ACP^C{}_1$ in tribal people is for the Kadar of Kerala[22] where two examples of the phenotype ACP_1 BC were detected.

6-Phosphogluconate dehydrogenase :

Indian populations generally have low frequencies of the common variant allele PGD^C, though the Kadar and the Malayarayans from Kerala have frequencies for this allele of 16.8% and 11.2% respectively[22 23]. In Andhra Pradesh the Savara and Jatapu from Srikakulum District were monomorphic for PGD^A [17] whilst the Kolams from Adilabad District[15] and the Chenchu from Mahabubnagar and Kurnool Districts[16] had frequencies for PGD^C in the vicinity of 2%. The values for PGD^C in the six tribal populations reported in this paper are therefore within the range for this allele in tribal populations in Andhra Pradesh in particular, but also for Indian populations in general.

Adenylate kinase

The frequency of the allele, AK^2, is generally of the order of 10-12% in Indian populations. In Andhra Pradesh, Santachiara-Benerecetti *et al*[25]

have reported AK^2 frequencies of 6.6% for the Koya Dora and 7.4% for the Konda Reddis. Rao *et al*[17] have reported 12.1% for the Savara and 3.2% for the Jatapu. Ramesh *et al*[15] have reported 5.4% for Kolams and Ramesh *et al*[16] gave values of 8.3% and 5.7% for AK^2 in two Chenchu groups. It therefore seems that AK^2 gene frequencies for tribal people, at least in Andhra Pradesh, tend to be rather lower than for Indian populations in general. Thus, for the populations reported here, the frequencies of AK^2 are consistent with those for the other tribal groups mentioned above, though the value of 13.1% for the small sample of Sugali is rather high.

Phosphohexose Isomerase :

The literature now contains a number of references to the occurrence of PHI variants in Andhra tribal populations[15–17]; in the populations reported in this study, a single example of PHI 3-1 was detected in the Rajgonds and six examples in the Koya.

Genetic distance

Using Morton's[11] hybridity coefficient (θ) and Nei's[13] standard genetic distance measure (D), we have computed the matrices of genetic distances, Table IV, among 11 major tribal populations from Andhra Pradesh. The populations analysed are the six tribal groups in the present study, two Chenchu populations taken from Ramesh *et al*[16], the Kolams[15] and the Savara and Jatapu.[17] The calculations are based on data from 13 genetic (polymorphic) systems. These include ABO, MN, Rh, phosphohexose isomerase (PHI), lactate dehydrogenase locus B (LDH_B), adenylate kinase (AK), 6-phosphogluconate dehydrogenase (PGD), acid phosphatase-1 (ACP_1), phosphoglucomutase locus-1 (PGM_1), haemoglobin-β ($HB\beta$), haptoglobin (Hp), transferrin (Tf) and group specific component (Gc). The gaps in the data from our laboratory have been variously filled from published sources. Systems for which information on individual populations are lacking in the literature, have been given the default gene frequencies, computed as weighted averages from data on other Andhra tribal groups.

The matrix of hybridity coefficients (θ) has been reduced to construct a dendrogram (Fig. 2). The phenetic relationships demonstrated by this approach match well in parts, with the geographical proximity of the populations in question. For example, the two populations from the coastal district of Srikakulum, the Savara and Jatapu cluster together. Similarly, the Kolams, Koyas and Rajgonds from the Telangana districts of Khamman and Adilabad are grouped together. These two

neighbouring groups of tribes also cluster with each other. The four populations from the southern districts, the Mahabubnagar and Kurnool Chenchu, the Yerukulas and the Lambadi, however, do not show any such association and join the clusterings at random.

The two dimensional representation of the genetic distances, based on a plot between first two scaled eigenvectors (Fig. 3) gives a slightly different insight into the genetic relationships. The Pardhans show closeness with other populations from the Telangana region. The populations of Mahabubnagar and Kurnool Districts, however, still show random dispersion. Additional evidence, although subjective, obtained by reducing Nei's distance matrix (Table IV), however, reveals closer relationship between the Yerukula and Lambadi.

The pattern of clustering seen among the tribal populations of Andhra Pradesh, although consistent with their geographical positioning in general (these patterns are also consistent with their migration histories) has certain aberrations. Historically, some tribal groups, such as the Lambadi and Pardhans, are believed to be of north Indian origin and have migrated to these districts comparatively recently. The differences between the two Chenchu groups, on the other hand, have been accentuated by their small effective population sizes. Similarly, small sample size may be responsible for the inconsistent behaviour of the Yerukula.

The genetic distances computed by using Nei's method can be tested for their level of significance using standard errors. Two populations stand out in this regard : Chenchus (M) and Pardhans exhibit significant differences in genetic distances with each other as well as in 13 out of 18 possible pair-wise combinations with others. Lambadis are also quite deviant in this regard. Among non-significant values, the contribution from Chenchus (K) is the largest; the group does not show significant distance from any of the populations under consideration.

Conclusions

The conclusions arrived at in this study with respect to the population structure and spatial heterogeneity in Andhra tribes are at variance with the results obtained by Chakraborty and Yee[3] from phenotype bioassay of five tribes from the adjoining district of Koraput in Orissa. The mean amount of the kinship coefficient (ϕ) in Andhra tribes is 9.523×10^{-3} in comparison with 2.613×10^{-3} for Orissa tribes. Similarly, mean within population estimates of ϕ_{II} are much larger in Andhra tribes than in Orissa tribes (19.051×10^{-3} against 6.657×10^{-3}). These differences may be accounted for by the use of a different set of genetic markers with different spectra of allelic variability and also the basic difference in their mating

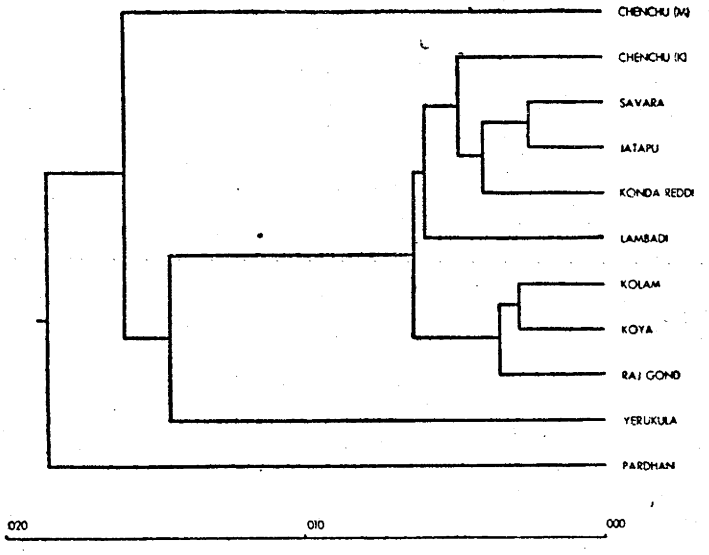

Fig. 2. Phenetic relationships among 11 Andhra tribal populations based on Morton's θ_{ij} values.

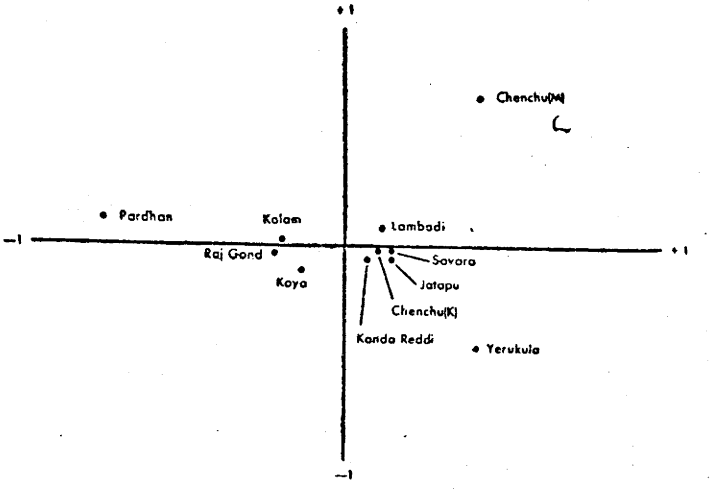

Fig. 3. Eigenvectorial representation of genetic distances based on Morton's hybridity coefficients, θ_{ij}.

patterns. The latter difference is significant in view of the very high proportion of consanguinity in Andhra tribes ($\hat{F}$=0.020–0.47)[27], and the formation of relatively small sized inbred isolates, by virtue of their thin dispersion, in Chenchus, Yerukulas, Lambadis and Pardhans. The Orissa populations on the other hand avoid consanguinity and show relatively continuous distributions.[3,4]

Another major difference with the analysis of Orissa tribes by Chakraborty and Yee[3] is the role of geographical proximity in influencing the mean value between populations of the hybridity coefficient, θ_{ij}. The most distant populations exhibit maximum amount of hybridity coefficient in Andhra, while this difference is low in adjacent tribes or in tribes from the same area. The congruence of geographical placements with the topology derived from the θ_{ij} matrix is an indicator of the role of isolation by distance. This is not to suggest, however, that initial fissioning and settlement of these populations does not affect this hierarchial classification. These results suggest a diametrically opposite situation in Andhra tribes from Orissa tribes.

The divergent placement of the two Chenchu groups on the dendrogram (Fig. 2) and the two dimensional representation (Fig. 3), however, is unexpected. Loci contributing heavily to the high values of θ between these two groups are the phosphohexose isomerase (PHI) and the lactate dehydrogenase locus (LDH_B), both of which are polymorphic in the Chenchus from Mahabubnagar but show no variation in the Chenchus from Kurnool[16]. A new matrix of estimates of θ_{ij}, calculated after excluding these two loci, not only brings the two Chenchu groups close together, but also demonstrates their close affinity with other autochthonous tribes of Andhra Pradesh. It demonstrates also the importance for genetic distance estimates of random drift effects on the frequency of alleles in small groups as exemplified here by the two Chenchu samples.

Acknowledgements

This work was carried out under the UGC Project, Anthropogenetic Studies of Andhra Pradesh Tribes (J.S. Murty). We wish to acknowledge the support and advice given by Professor O.S. Reddi in Hyderabad and Dr R.L. Kirk in Canberra.

References

1. Blake NM, Omoto K, Kirk RL and Gajdusek DC (1973) Variation in red cell enzyme groups among populations of the Western Caroline Islands, Micronesia. Am J Hum Genet *25* : 413-421.

2. Blake NM and Omoto K (1975) Phosphoglucomutase types in the Asian-Pacific area : a critical review including new phenotypes. Ann Hum Genet (Lond.) *38* : 251-273.
3. Chakraborty R and Yee S (1973) Phenotypic bioassay of five tribes of Orissa, India. Hum Hered *23* : 270-279.
4. Das SR, Mukherjee DP and Sastry DB (1968) A somatological survey of five tribes in the Koraput District, Orissa. Bull Anthrop Survey of India *17* : 400-422.
5. De Jong WW, Bernini LF and Meera Khan P (1971) Haemoglobin Rampa : α95 Pro→Ser. Biochim Biophys Acta *236* : 197-200.
6. De Jong WW, Meera Khan P and Bernini LF (1975) Haemoglobin Koya Dora : high frequency of a chain termination mutant. Am J Hum Genet *27* : 81-90.
7. Ghosh AK, Kirk RL, Joshi SR and Bhatia HM (1977) A population genetic study of the Kota in the Nilgiri Hills, South India. Hum Hered *27* : 225-241.
8. Goud JD. *Population genetics of five endogamous tribal groups of Andhra Pradesh.* Ph.D. Thesis. Osmania University, Hyderabad, India, 1977.
9. Goud JD and Rao PR (1977) Distribution of some genetic markers in the Yerukala tribe of Andhra Pradesh. J Ind Anthrop Soc *12* : 258-265.
10. Goud JD and Rao PR (1980) Transferrin, haptoglobin and group-specific component types in tribal populations of Andhra Pradesh. Hum Hered *30* : 7-11.
11. Morton NE. Kinship information and biological distance. In : Morton NE, ed. *Genetic structure of populations.* Honolulu : Univ of Hawaii Press, 1973.
12. Naidu JM and Veerraju P (1977) A_1 A_2 BO and Rh (D) blood groups among tribals of Andhra Pradesh, India. Acta Anthropogenet *1* : 36-40.
13. Nei M (1972) Genetic distance between populations. Am Naturalist *106* : 283-292.
14. Ramachandraiah T. *A genetic study of the Lambadis of Andhra Pradesh.* Ph.D. thesis. University of Delhi, India, 1967.
15. Ramesh A, Murty JS and Blake NM (1979) Genetic studies on the Kolams of Andhra Pradesh, India. Hum Hered *29* : 147-153.
16. Ramesh A, Blake NM, Vijayakumar M and Murty JS (1980) Genetic studies on the Chenchu Tribe of Andhra Pradesh, India. Hum Hered *30* : 291-298.
17. Rao PM, Blake NM and Veerraju P (1978) Genetic studies on the Savara and Jatapu tribes of Andhra Pradesh, India. Hum Hered *28* : 122-131.
18. Rao PR and Goud JD (1979) Sickle-cell haemoglobin and glucose-6-phosphate dehydrogenase deficiency in tribal populations of Andhra Pradesh. Indian J Med Res *70* : 807-813.
19. Rao TV and Reddy GG (1973) A population genetic study of the Bagatha and Valmiki tribes of Visakhapatnam District (A.P.), India. Jap J Hum Genet *18* : 226-236.
20. Rao VVR (1977) ABO blood group frequencies among six tribal communities in East Godavari. Man in India *57* : 127-136.
21. Saha N and Kirk RL (1973) A simple method for collecting blood for population studies. Hum Hered *23* : 182-188.
22. Saha N, Kirk RL, Shanbhag Shaila, Joshi SR and Bhatia HM (1974) Genetic studies among the Kadar of Kerala. Hum Hered *24* : 198-218.
23. Saha N, Kirk RL, Shanbhag Shaila, Joshi SR and Bhatia HM (1976) Population genetic studies in Kerala and the Nilgiris (South west India). Hum Hered *26* : 175-197.
24. Santachiara-Benerecetti SA, Cattaneo A and Meera Khan P (1972) Rare phenotypes of the PGM_1 and PGM_2 loci and a new PGM_2 variant allele in the Indians. Am

J Hum Genet *24*: 680-685.
25. Santachiara-Benerecetti SA, Cattaneo A and Meera Khan P (1972) A new variant allele AK^5 of the red cell adenylatekinase polymorphism in a non-tribal Indian population. Hum Hered *22*: 171-173.
26. Simmons R, Graydon J, Semple N and D'Sena G (1953) A genetical survey in Chenchu, south India : blood, taste and secretion. Med J Aust *i* : 497-503.
27. Veerraju P. Consanguinity in tribal populations of Andhra Pradesh. In : Verma IC, ed. *Medical Genetics in India*, Vol. 2. Pondicherry : Auroma Enterprises, 1978; 157-163.

Dr N. M. Blake
John Curtin School of Medical Research
P. O. Box 334
Canberra, A.C.T. 2601
Australia