PROGRESS REPORT

AEC CONTRACT AT(11-1)-1552

AREA PROGRAM IN POPULATION GENETICS

NOTICE

This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the United States Atomic Energy Commission, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represents that its use would not infringe privately owned rights.

Department of Human Genetics
University of Michigan Medical School
Ann Arbor, Michigan

MASTER

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

V. PROGRESS REPORT

The progress made during this past year can best be summarized under five headings.

V.1. THE GENETIC PARAMETERS OF AMERINDIANS, WITH
PARTICULAR REFERENCE TO THE YANOMAMA

*Appendices
removed
reprints
fragments*

A vigorous effort has been made the past year to summarize many aspects of the work in progress over the past 10 years. Where a paper has been published during the past year, or where a manuscript is in press, we propose only to indicate the major findings, and attach the paper or manuscript as an Appendix. Where, however, the treatment of a subject is not yet so finalized, we propose to supply a short summary. It is appropriate to point out that Neel is committed to the preparation, during 1974-1975, of a volume (based largely on reprints) summarizing all this work, to be one in the series of volumes reviewing the results of the International Biological Program.

V.1.1. Comparisons of Different Measures of Biological Distance Between Populations (Based on Genetic Markers, Anthropometrics, Dermatoglyphics and Geography) (da Rocha, Lingoos, Neel, Rothhammer, Smouse, Spielman). - In order to determine the degree of correspondence between sets of multivariate observations based on different kinds of traits (genetic markers, anthropometrics, and geographic distance), Spielman (1973) (Appendix 1) developed two new statistical methods, derived from different notions of "correspondence" and compared the results of an analysis of seven Yanomama villages. The first method tests the similarity of the cluster or hierarchic structures implicit in two sets of data, while the second tests the departure from perfect geometric congruence or superimposability. Although both methods indicated significant correspondence between some of the sets of data, the direct test of congruence gave results which differed substantially from those based on cluster structure.

Neel, Rothhammer, and Lingoos (1974) (Appendix 2) then extended this type of comparison by adding dermatoglyphics to the three

kinds of data employed by Spielman (1973). They have also extended the statistical treatment to include the use of non-parametric techniques. Using a different (but overlapping) set of 7 villages from that employed by Spielman, they generally observe more significant agreement between the picture of internal diversification provided by the various approaches than in Spielman's analysis. This is somewhat surprising for the case of dermatoglyphics, in whose etiology genetic factors are thought to be less important than for genetic markers. The suggestion is made that for tribal-type populations the principal methods currently in vogue for examining the internal relations of population sub-units are yielding more or less equivalent results, so that choice of technique in these circumstances becomes largely a matter of individual preference rather than some statistical advantage.

The foregoing treatments have involved intratribal subdivisions (Yanomama villages). Da Rocha, Spielman, and Neel (in press)(Appendix 3) have extended the treatment to a mixture of intra- and intertribal units (seven villages drawn from four different tribes). Only genetic marker and anthropometric data were available for this purpose. The "distances" between genetic units were now greater on the average than in the preceding treatments, and the measures of congruence between distance matrices or dendrographic representations significant at even lower P-levels.

On the basis of the foregoing experience, Spielman and Smouse have progressed to a related and somewhat more involved problem. They have investigated the question of structural similarity by multivariate canonical correlation analysis, but with data from 19 villages. For the 6 codominant allele pairs which constitute the genetic variables, individual genotypes were converted to the "measurements" 0, 1, or 2, i.e., a count of the individual's alleles of one type. The anthropometric data are continuous; some of the 12 variables were transformed to eliminate skew and all were adjusted for age. A conventional among-groups covariance matrix was computed, explicitly incorporating non-zero covariances between genetic variables.

The technique of canonical correlation finds one linear combination (x_1) of the A variables and another (y_1) of the G variables, chosen so that the correlation of the two linear indices is maximized; this is the first canonical correlation. Next, a second combination of the A variables is formed, as is another for the G variables. Subject to the constraint

that both are uncorrelated with x_1 and y_1 , these are chosen so that their correlation is maximized; this is the second canonical correlation. There are as many such canonical correlations as in the smaller set of variables. For our data, the six canonical correlations are: (.915, .893, .815, .693, .372, .304). The first two of these are significant ($p < .05$), indicating a high degree of association between the village means for the A and G variables considered as a whole; this is equivalent to saying that villages occupy the same relative positions in the A-space as in the G-space.

Non-random association in structure implies differentiation of village means, which in turn suggests ability to discriminate among villages; given an individual's 18 measurements, it should be possible to specify with better-than-chance success which village he belongs to. This has led us to ask: How accurately can we assign individuals to the correct village (represented by the village means), using A alone, using G alone, and using the two kinds of traits together? Data for men and women from the same village were combined after standardization of A traits in each sex to mean zero and unit variance. For each individual, the generalized distance (Mahalanobis' D^2) to all 19 villages was computed, and the individual was assigned to that village yielding the smallest distance.

The twelve A traits alone give correct village classification of 46% of the 520 individuals studied. Of the erroneous assignments, more than a third are to villages closely related historically to the correct village. Using these historical and geographic "neighborhoods" to define nine clusters of villages, we find that 66% of the individuals are assigned to the correct cluster. The six G traits yield correct village classification of only 19%, while only 30% are assigned to the proper cluster. Table 1 shows these results, as well as the results for classification based on the combination of A and G traits. It is apparent that the full battery of 18 measurements is a powerful tool for classification, and accordingly that considerable differentiation of villages and clusters exists.

Is the marked advantage of classification by A compared to G due simply to the larger number of traits? The last line of the table shows the mean percentages obtained with two non-overlapping selections of only six anthropometric traits. Even with these reduced sets of A variables, classification is 50% better than with the G traits. These results are

Table 1. Recent correct classification of 520 Yanomama Indians by individual measurements and village means for 19 villages, constituting 9 clusters of related villages. For evaluation of significance of the correct classification, the 10 villages were constructed artificially by drawing 520 individuals randomly (with replacement) from the total. The percentage of correct assignments to these "random" villages was determined, and the process was repeated 5 times, giving the means and standard deviations shown.

	Correctly assigned to		Correctly assigned to cluster	
	Real villages	"Random" villages	Real villages	"Random" villages
Anthropometric traits (12)	46	11 $\pm$ 1	66	19 $\pm$ 3
Genetic traits (6)	19	10 $\pm$ 1	30	16 $\pm$ 3
Combination (18)	55	14 $\pm$ 2	71	21 $\pm$ 2
Anthropometric traits (6)	30		47	

preliminary and classification by Mahalanobis' D^2 may not be optimal for 3 traits. Nevertheless, the difference between classification accuracy using A and G traits is large enough to suggest that the two sets of traits are subject to significantly different kinds of differentiation. Since the anthropometric traits are influenced by environment while the strictly genetic traits are not, these observations suggest further attempts to identify regional differences in environment which might affect the development of adult morphology.

V.1.2. The Contribution of Size and Shape Differences to Village and Sex Differences in Body Measurements (Spielman). - In this study Spielman (1973) (Appendix 4) presented a new technique which partitions the morphological differences between men and women of the same or different villages into a component of size difference and a component of shape difference. The analysis demonstrates that once gross size differences have been partitioned out, men and women from the same village are more similar morphologically (i.e., in shape) than men and women from different villages. This relative homogeneity in morphology within groups is presumably the counterpart of the genetic microdifferentiation among groups, abundantly demonstrated elsewhere.

V.1.3. The Occurrence of Rare ("Private") Variants of 18 Different Serum Proteins or Erythrocyte Enzymes in Six South American Indian Tribes (Tanis, Neel, Dovey, Morrow). - Four of the group (1973) (Appendix 5) have increased the data available on the occurrence of rare variants in Indian populations, reporting at least nine new variants in a series of 31,806 system-determinations.

V.1.4. Estimates of the Rate of Mutation Resulting in Electrophoretically Distinguishable Variants in Serum Proteins and Erythrocyte Enzymes. - Neel (1973) (Appendix 6) has subjected the data described in V.1.3 plus previous similar data, amounting in all to 56,237 system determinations, to a formulation developed by Kimura and Ohta (1969). Given data on mean survival time for a selectively neutral mutant and the size of the breeding population, this formulation permits estimation of the mutation rate responsible for the observed variation. The formulation is

$$\mu = \frac{I}{2N} \cdot \frac{1}{\bar{t}_0}$$

where I = the average number of mutant alleles per locus among the loci sampled, N = the number of individuals in one generation in the population in which I has been determined, and $\bar{t}_0$ = the average mutant survival time in generations for those mutants not going to fixation. The development of a figure for mean mutant survival time by means of computer simulation of an Indian population was reported last year. By definition, I should not include those variants that will ultimately go to fixation, but this presumably very small fraction of the total cannot be excluded from any actual data set. The final formula has an appealing simplicity: on the assumption of equilibrium, the mutation rate is equal to the mean proportion of different variants per locus ($I/2N$) times the fraction of these variants lost each generation ($1/\bar{t}_0$). While the assumption of neutrality of mutant phenotype is essential to this formulation, it is clear that the effect of the most probable departure from this assumption (deleterious mutants) will be an underestimate. The estimate for mutations resulting in electrophoretically detectable variants is 3×10^{-5} /locus/generation and for both detectable and electrophoretically silent, 8×10^{-5} /locus/generation.

Since the unit of study for these investigations of Indian populations is the village, whenever a rare variant is encountered, family data are usually available. Thus, if both the father and mother of a person with a rare variant are included in the sample, one has a direct test of whether a variant has originated by mutation in the preceding generation. To date, these trios provide a total of 30,480 "locus tests," in which the occurrence of an allele in a child not present in either parent could have been detected. Thus far there is no evidence for mutation. Since expectation based on the estimate of the preceding paragraph was to detect only one mutation, there is as yet no significant conflict between prediction and observation.

V.1.5. The Numbers and Frequency of Histocompatibility Types and Alleles in the Yanomama (Layrisse, Layrisse, Malavé, Terasaki, Ward, Neel). - Two of this group with four collaborators (1973, Appendix 7) have reported the occurrence of at least 5 alleles of the first HL-A series and six of the second series among the Yanomama. It is suggested that the maintenance of this degree of genetic diversity, if indeed only two genetic loci are

concerned with the determination of these specificities, strongly suggests selection on the phenotypes associated with these alleles.

V.1.6. The Uniqueness of the Yanomama. - Ward, Gershowitz, Layrisse, and Neel (in press)(Appendix 7) demonstrate that the Yanomama appear to be genetically distinct from 20 other South American tribes in their distribution of allelic frequencies for the following 6 loci: Rh, MNS, Fy, Jk, Di and Hp. A consideration of their greater genetic distinctiveness from geographically adjacent tribes compared with more distant tribes, and their relatively low levels of heterozygosity reinforces the impression they have long been a genetically isolated population. This conclusion is also reached by considering the over-all relationship between a tribe's genetic distance from the inter-tribal average gene frequencies and the geographic location of that tribe: Whereas the majority of the 20 tribes display a consistent relationship of increasing genetic disparity with increasing distance from a central "corridor" running into the Amazon Basin from the Isthmus, the Yanomama stand out as quite disparate. Fig. 1 presents this finding graphically as a contour diagram. Presentation of tribal gene frequencies in this fashion brings an apparent "regularity" into tribal relationships, a greater regularity than is seen in dendrographic representations. However, the interpretation of this representation will undoubtedly be the subject of much discussion. One attractive possibility is that the long axis of this plot represents the principal route of advance of the Amerindian into South America, with the gene frequencies of tribes in and near this corridor most nearly corresponding to the "original" frequencies of the Indians entering South America. A very different interpretation is that the Indian groups entering South America some 20,000 years ago went off in all possible directions, but that subsequent fusions of these various groups in the central portion of South America led to the gene frequencies there being the closest to the over-all average.

V.1.7. Linguistic and Genetic Differences Among Yanomama Language Areas (Spielman, Migliazza, Neel). - The extent of linguistic differentiation and its correspondence to genetic differentiation have been examined in seven language areas of the Yanomama (Appendix 9). Using one of the

Figure 1. An isocline plot of the genetic distance of 20 South American Indian tribes from the mean of all South American tribes, superimposed on a map of South America. As is customary in such plots, the centroid has been located to conform to the actual genetic distances from the mean of all tribes, which mean is based on the best estimate of the gene frequencies of the original population entering South America. Each contour corresponds to an increment of 0.03 genetic distance units, the central contour embracing a distance of 0.205 from the mean of all tribes.

objective methods mentioned earlier for quantifying the degree of similarity in the clustering or hierarchic relationships implicit in different kinds of data, the linguistic divergence of the seven areas is shown to correspond significantly with the pattern of genetic differentiation. This conclusion is corroborated by another of the tests developed by Spielman for measuring geometric congruence of the patterns of village positions specified by two different kinds of data. The analogy between linguistic and genetic divergence, first recognized well before 1900, is redeveloped in the light of more recent population genetics.

Approximate time depths, necessary for an estimate of the rate of genetic divergence, have been calculated from cognate counts among the dialects. The time depths suggest that the Yanomama dialects have been diverging for about 1000 years. Cognate percentages with non-Yanomama languages lead to the tentative conclusion that Yanomama languages as a whole diverged at least 1000 years earlier still. These linguistic findings are consistent with the well-documented genetic finding that the Yanomama show a high degree of microdifferentiation in allele frequencies within the tribe, with a marked genetic distinctiveness when compared to other South American tribes.

V.1.8. The Demography of the Yanomama (Neel, Weiss). - Over the years, considerable data have been accumulated on the demographic balance of the Yanomama, more such data than has heretofore been available for any population of this type. We (in press, Appendix 10) have now summarized this material, deriving life tables, fertility schedules, and estimates of gene flow not reflected in pedigrees (non-paternity). The treatment is seen as providing the basis for more realistic demographic models for primitive populations than have existed in the past. In addition, the data provide for relatively precise estimates of such demographic measures as Fisher's Reproductive Value, Crow's Index of Potential Selection, and Weiss' Index of Growth Regulation.

V.1.9. A "Private" Genetic Polymorphism (Tanis, Ferrell, Neel, Morrow). - The paradoxical term "private" genetic polymorphism is coined to describe the presence of a polymorphism for an albumin variant in the Yanomama, in a gene frequency of 0.08, which has not been found (with one dubious

exception) in any member of the 9 other tribes examined to date (Tanis, Ferrell, Neel, and Morrow, in press, Appendix 11). The polymorphism has the additional unusual feature that within the Yanomama its distribution exhibits a significant geographic cline. The manuscript describing this situation also contrasts the relative frequency and variety of albumin variants in the Amerindian with the paucity of hemoglobin variants, and conclude that the interplay of selection and mutation are probably different for the respective genetic loci.

V.1.10. A Comparison of Various Analyses of Population Structure (Spielman, Neel). - Subdivision of a population, which may result from demographic and geographic factors, leads to nonrandom association of alleles for a given locus. Wright (1965) defined three correlations (F-statistics) for the association between alleles in single or different (random) individuals, and showed how these F-statistics are related to each other and to types of subdivision of the total population. F_{ST} (which we also call θ) is the correlation between random gametes within subdivisions. F_{IS} is the correlation between alleles in a single individual within a subdivision, averaged over subdivisions. F_{IT} is also a correlation between alleles in an individual, but relative to the total population. The relationship between these correlations, taken as parameters, is: $F_{IT} = F_{ST} + (1 - F_{ST})F_{IS}$.

Viewed as intraclass correlations, the F-statistics are ratios of within-group covariances to within-group variances, therefore equivalent to ratios of variances. This led C. C. Cockerham to make explicit the analogy with variance components analysis. The occurrence of alleles in individuals, of individuals within subdivisions, and of subdivisions within the total population defines a nested analysis of variance, which Cockerham has developed, with corresponding estimates of F-statistics as ratios of variance components which are therefore least squares estimates. The subdivision of the Yanomama population into villages offers a natural application of this technique, and a testing-ground for the estimation of structure parameters generally.

In the preliminary analysis completed at the time of this report, we follow two strategies. First we fix the estimation technique

(least squares) and vary the subsets of data used (partitioning our current set of data on 48 villages by location or sex). Here we ask: How do the F -statistics change with partitioning of the data? The second strategy is to fix the body of data [all individuals tested in 37 villages studied by Neel and Ward (1972)] but vary the technique of estimation. The question is: Do different methods which ostensibly analyze the same structure give different answers with a single body of data?

Table 2 gives the least squares estimates for all 48 villages now available with sample size greater than 30. As estimated here, F_{IT} and θ are about .05 and .06 respectively. On the assumption that the mean squares in the ANOVA are distributed as chi-square, the differences among villages are a highly significant source of variation; this is equivalent to finding θ or F_{ST} significantly different from zero, and is observed for every locus. However, in no case were differences among individuals within villages a significant source of variation; this is equivalent to finding no significant excess of F_{IT} over θ , or ultimately, no significant F_{IS} .

One feature of structure which depresses θ and therefore F_{IT} is migration or gene flow among villages. To reduce the importance of migration in the units studied, we selected eleven villages which are highly dispersed geographically and known from anthropological data not to have substantial interchange. This is the first of the partitions of the data. The results, given in the second line of Table 2, show that F_{IT} and θ , while both slightly larger and more similar than before partitioning are not greatly changed. Since it is known that very little migration is required to eliminate the effects of subdivision, perhaps we have failed to obtain truly isolated units by our rather ad hoc treatment. Thus we have not obtained evidence that this partition of the data significantly affects the F -statistics.

Anthropological and demographic observations suggest that men and women make differently structured genetic contributions to the next generation; in a given village, married women are more likely to be immigrants than are married men. The result of partitioning the data by sex is shown in the third line of Table 2. For the men, θ or F_{ST} is larger than F_{IT} , leading to a negative F_{IS} (excess of heterozygotes). For the

Table 2. Least squares estimates (intraclass correlations) of structure parameters. Entries are means $\pm$ standard deviation for 8 allele pairs (M/N, S/s, C/c, E/e, Fy^a/Fy^b, Hp¹/Hp², Gc¹/Gc², PGM¹/PGM²).

		F_{IT}	$\theta = F_{ST}$	F_{IS}
Men and women combined	48 villages	.051 $\pm$.034	.062 $\pm$.016	-.012 $\pm$.025
	11 geographically dispersed villages	.066 $\pm$.047	.065 $\pm$.029	.000 $\pm$.038
Men only	48 villages	.052 $\pm$.038	.072 $\pm$.020	-.021 $\pm$.037
Women only	48 villages	.059 $\pm$.046	.059 $\pm$.016	-.001 $\pm$.042

women, however, F_{IT} and θ do not differ appreciably, so that F_{IS} for them is close to zero. Since the women in a village represent an amalgamation drawn from various other villages, their genotype frequencies should show an excess of homozygotes, relative to those for men within villages. While the difference in F_{IS} between the sexes may not be statistically significant, F_{IS} is more positive in women than in men; the difference is in the direction expected on theoretical grounds as a consequence of social structure.

Finally we consider varying the technique of estimation. This is not merely a methodological issue; substantive considerations appear in this question, as well, since ability to detect locus differences in structure is important in a test for selection. Rothman, Sing, and Templeton (in press) have developed a maximum-likelihood (ML) technique; the method they use has, as a parameter, the correlation (ρ) between genotypes within villages. Estimation consists in finding that value of ρ which maximizes the likelihood of the observed array of genotypes. Table 3 shows estimates of (ρ) obtained by three different methods. Using the data of Neel and Ward (1972) for comparability with their results, we see that their method and least squares estimation give similar estimates. However, the value given by ML is only about 60% of the other two. In fact, Rothman, Sing and Templeton have shown by simulation with known parametric ρ that although the ML estimate is biased (consistently smaller than the true value), the mean squared error (deviation) of ML estimates about this true value is less than that for the least squares estimate by a factor of two. Since a real estimation situation permits only one estimate, the method with smaller mean squared error is preferable. The next step is to use ML procedures to test the heterogeneity of structure parameters for the loci examined here.

V.1.11. What Is the Pattern of Gene Flow Within a Tribal Population? (Ward, Neel). - The term "gene flow" is readily used in population genetics but its "grass root" dynamics are poorly understood. By chance, the Yanomama appear to be an unusually favorable group for the study of this phenomenon. On the one hand, as noted above, they possess a unique polymorphism (albumin Yanomama-2) which documents outward gene flow. On

Table 3. Comparison of three estimates of Wright's coefficient of relationship (Rho), the correlation between genotypes within subdivisions, for 37 Yanomama villages studied by Neel and Ward (1972). Entries are means for 8 loci $\pm$ standard deviation.

	Mean rho $\pm$ std. dev.	Estimated by:
Least squares	0.110 $\pm$.032	$2\theta/(1 + F_{IT})$
Neel and Ward (1972)	0.124 $\pm$.029	$2F_{ST}/(1 + F_{IT})^*$
Maximum likelihood	0.070 $\pm$.022	Search of likelihood surface

*For each locus: $F_{ST} = \sigma_p^2/\bar{p}\bar{q}$; σ_p^2 is weighted sum of squares of village deviations from mean allele frequency $\bar{p}$, $\bar{q} = 1 - \bar{p}$.

$F_{IT} = 1 - \frac{H_T}{2\bar{p}\bar{q}N_T}$; H_T is total number heterozygotes, N_T is total number of individuals tested.

the other hand, they appear to have lacked until recently two other polymorphisms commonly found in the surrounding tribes (the Diego and acid phosphatase polymorphisms). We have studied the process of admixture as it appears to have occurred in the Yanomama during the past century with particular reference to two attributes of such admixture events: the introduction of new alleles and differential rate of incorporation of migrant alleles as a function of socio-cultural practices. In addition, there are clines of gene frequency which probably cannot be wholly explained by admixture and may in part result from recent population expansion.

V.1.11.1. Introduction of New Alleles. - The influence of admixture in introducing new alleles into the Yanomama can be discerned in two areas. In the Northeastern sector, three villages possess both the Di^a and Ph^A alleles in appreciable frequencies, as a consequence of a series of historically documented admixture events involving the Makiritare, Maku and Awake' tribes. The Di^a allele is only found in this area, its presence being attributable to gene flow from adjacent Carib tribes to the North. The Ph^A allele appears to have been also introduced in the Southwest, as it is present in 12 villages scattered along the southwestern periphery of the Yanomama. The majority of these unrelated villages lie outside the tribal boundaries of a century ago and apparently possess the Ph^A allele as a consequence of a series of sporadic contacts with Arawak groups as the Yanomama intruded into this region. Since this complex of Arawak tribes (Kuriobana-Barauana) in all probability lacked the Di^a allele (in common with other Arawak groups to the West) the absence of Di^a from this region is readily explicable. The latter series of admixture events involving the Arawak groups in the Southwest resulted in a widely dispersed distribution of the Ph^A allele in this area, with some indication from the Ocamo region that it is beginning to diffuse back into the rest of the tribe.

V.1.11.2. Rate of Incorporation of Migrant Alleles. - In subdivided populations, such as the Yanomama, the rate of flow of migrant genes across tribal boundaries is largely a consequence of the rate of incorporation of alleles into the deme participating in the admixture event. There are two observations indicating that genes are more easily incorporated into the Yanomama than other tribes. On one hand, there is the virtual

absence of the Albumin^{Yan-2} allele from adjacent tribes compared with the ready introduction of the Di^a and Ph^A alleles into the Yanomama. The lack of flow of the Albumin^{Yan-2} allele is even more remarkable in the face of 150 years contact that the Makiritare have had with the Sanema, the Yanomama group possessing the highest frequency of this endemic polymorphism. On the other hand an analysis of pedigree data for two Makiritare villages and two Yanomama villages, indicates not only that the expected proportions of admixture are much higher for the Yanomama (average of 30% vs 10%) but that the effective migration rate of the foreign alleles increases in successive generations in the Yanomama villages while it decreases, or at best remain stable in their Makiritare counterparts. The enhanced flow of genes into the Yanomama is a consequence of socio-cultural factors which act to give migrants and their descendants a greater than average fertility in Yanomama villages, the reverse being observed in Makiritare villages.

V.1.11.3. Geographic Structure. - One consequence of admixture is to enhance the degree of geographic structure in the subdivided population, especially when coupled with a lineal expansion of subgroups as in the case of the Yanomama. Geographic structure was investigated by use of the clinal regression techniques of Smouse (1974) which facilitates partitioning of the components of regression and testing for goodness-of-fit. All 9 of the codominant alleles tested and all 5 of the dominant alleles display a significant regression (clinal) component ($p < .05$). Since the residual (lack of fit) component is also significant in each case, the goodness-of-fit is obtained by taking the ratio of these two components, each divided by the appropriate degrees of freedom. The probability levels for goodness-of-fit, assuming an asymptotic F distribution, are designated as follows in Table 9: * $p < .05$; ** $p < .01$. Alternatively stated, of 14 independent alleles (counting M-S and C-E separately) 7 have one or more alleles displaying a cline which accounts for a greater portion of the variance in village gene frequencies than the residual.

Since the distribution of Di^a and Ph^A alleles indicates that admixture occurred in the Northeast and Southeast, alleles with significant clinal variability along this axis (M, Gc¹, Ph^A) may owe their present distribution to admixture events and the subsequent flow of genes. Alleles whose clinal variability is predominantly along the Northeast-Southwest axis (C, Yan-2, Jk^a, Gm^{axg}) may owe their geographic structure more to the effects of internal

Table 4. Analysis for the occurrence of clines in the Yanomama. All 9 of the codominant alleles tested and all 5 of the dominant alleles display a significant regression (clinal) component ($p < .05$). Since the residual (lack of fit) component is also significant in each case, the goodness-of-fit is obtained by taking the ratio of these two components, each divided by the appropriate degrees of freedom. The probability levels for goodness-of-fit, assuming an asymptotic F distribution are designated as follows: * $p < .05$; ** $p < .01$.

Quadrant		Alleles
a) <u>Codominant alleles</u>	Frequency increases from Northwest to Southeast	E
	Frequency increases from Southeast to Northwest	C**; Yan-2**; Hp ¹ ; PGM ₁ ¹
	Frequency increases from Southeast to Northeast	Ph ^{A**} ; S
	Frequency increases from Northwest to Southeast	M*; Gc ^{1*}
b) <u>Dominant alleles</u>	Frequency increases from Northwest to Southeast	Gm ^{axg**}
	Frequency increases from Southeast to Northwest	Jk ^{a**}
	Frequency increases from Southeast to Northeast	P; Le ^a
	Frequency increases from Northeast to Southwest	Fy ^a

expansion than gene flow. Since, however, for the latter alleles, admixture cannot very well be invoked as a cause of the clines, it is quite possible it should not be invoked as a cause for the clines in the former alleles. Analyses of the extent of the admixture which would have to be postulated are underway. The extent of geographic structure in allelic distribution of the Yanomama is sufficient indication of the desirability of obtaining genetic information on the Arawak tribes to the South in order to evaluate the role played by gene flow in influencing the magnitude and distribution of genetic microdifferentiation in a subdivided tribal population.

V.1.12. The Gm and Inv Allotypes (Gershowitz). - The South American Indian populations which we have typed and reported (Yanomama, Makiritare, Wapishana, Macushi and Piaroa) are characterized by very high frequencies of Inv¹ (.25 - .75) and by the relative absence of certain Gm haplotypes which might be expected to be found in Mongoloid or Mongoloid-derived populations.

The Yanomama are unique, in that only two Gm haplotypes, Gm^{ag} and Gm^{axg} have been found in the entire population. As noted for other polymorphic systems, there is extensive microdifferentiation, the village gene frequencies for Gm^{axg} varying from .01 to .46, although the gene frequency estimated for the entire population is .15. The reason for the failure to find any evidence for the existence of two characteristically Mongoloid haplotypes, Gm^{afb} and Gm^{ab⁰st}, is still not readily apparent.

Another feature of the Gm haplotypes among the Yanomama is a significant clinal distribution. The Gm^{axg} haplotype is found in highest frequency (.287) at the eastern edge of the Yanomama range and falls progressively to the southwest, to .113. The cline is clearly defined by the six geographic areas described by Neel et al. (1971) on the basis of "natural groupings" detected by a genetic network analysis.

Among the other tribes, other Gm genes have been detected, whose source is uncertain. A brief word of explanation should be interjected here to explain the cause of the uncertainty. The Gm specificity Gm(b¹) can be found in all races of man, but in different haplotype associations. In Caucasians, the (b¹)-containing haplotype is Gm^{fb}, in Mongoloids it is Gm^{afb}, and in Negroes it is Gm^{ab}. The existence of Negro-Caucasian admixture is detected by certain indicator phenotypes which can

only be due to heterozygosity for Negro/Caucasian haplotypes (e.g., the Gm(agb) phenotype is due to $\text{Gm}^{ab}/\text{Gm}^{ag}$, while Gm(afb) is due to $\text{Gm}^{ab}/\text{Gm}^{fb}$), since Gm^{fb} is not found in Negroes and Gm^{ab} is not found in Caucasians. Mongoloid-Caucasian admixture is not easily detected, however, since Gm^{fb} cannot be distinguished from Gm^{afb} in the presence of Gm^{ag} haplotypes, common in Mongoloids. Only the detection of an individual of the phenotype Gm(afb), genotype $\text{Gm}^{afb}/\text{Gm}^{afb}$ would prove the presence of this Mongoloid gene in an admixed population, and similarly, the detection of an individual of phenotype Gm(fb), genotype $\text{Gm}^{fb}/\text{Gm}^{fb}$ would be necessary to prove the existence of the Caucasian Gm^{fb} haplotype.

The Gm(b) specificity has been detected in low frequency in this study. Among 720 Makiritare, 4 persons were found to be of the phenotypes (agfb) or (axgfb). It is not possible to determine which genotype, $\text{Gm}^{a(x)g}/\text{Gm}^{fb}$ or $\text{Gm}^{a(x)g}/\text{Gm}^{afb}$, is involved. Suggestive evidence of Caucasian ancestry is obtained from the fact that all four probands are descended from a common Indian ancestor.

Among 186 Macushi, 19 persons were found who were non-Gm(ag) or non-Gm(axg); 18 alone were detected in village 26C. Eleven individuals, distributed among two families (further relationships unknown) were heterozygotes for the $\text{Gm}^{ab^{ost}}$ haplotype, a uniquely Mongoloid haplotype. But 7 additional persons were heterozygous for $\text{Gm}^{(a)fb}$. The seven were distributed among two sibships (further relationships unknown). The haplotype involved and the racial origin of the Gm(b) specificity cannot be deduced. Since both the Mongoloid Di^a blood group gene and the Mongoloid $\text{Gm}^{ab^{ost}}$ phenotype are also found in this village, one might suspect that the Gm(b) specificity is also Mongoloid in origin. One individual of the phenotype Gm(agb^{ost}) was detected in village 26B.

Among 62 Wapishana, two persons (father and son) heterozygous for the Negro Gm^{ab} haplotype were detected. There is documented evidence that Negro males have lived among the Wapishana since about 1835. Three persons (mother and two daughters) were $\text{Gm}^{(a)fb}$ heterozygotes and two persons (mother and son) were $\text{Gm}^{ab^{ost}}$ heterozygotes. But again, the source of the Gm(fb) specificities can only be inferred from the circumstantial evidence presented by known Mongoloid genes.

V.1.13. Blood Pressure, Salt Intake, and Renin and Aldosterone Levels in Yanomama (Oliver, Cohen, Neel). - In our early studies of the Yanomama, we observed that they failed to exhibit the progressive increase in blood pressure after age 20 characteristic of civilized populations. This led to an effort to evaluate sodium intake and serum and/or urine concentrations of the appropriate hormones. Oliver and Cohen find that the sodium excretion per 24 hours is of the order of 1 or 2 milliequivalents, an unbelievably low value by the standards developed from civilized populations. Conversely, potassium excretion is relatively high. Urine aldosterone levels are very high, but serum renin levels are less elevated. It would thus appear that in the face of chronic low sodium intakes, serum renin values are well below those seen in acute sodium deprivation, which had provided the serum renin values taken as the norm for low sodium diets.

V.2. PREPARATION FOR LARGE-SCALE MONITORING OF POPULATIONS FOR MUTATION RATES

For the past several years a substantial portion of the activities comprising this program have been predicated on the assumption that the study of both spontaneous and induced mutation in experimental animals and man was evolving toward the use of protein phenotypes. The activities summarized under V.1.4. represent an effort of this type. In this Section we briefly summarize two additional efforts in this direction.

V.2.1. The Pilot Study in Japan (Tanis, Ferrell, Neel). - Three years ago, Neel proposed to the Advisory Committee to the Atomic Bomb Casualty Commission (ABCC) of the National Research Council and also to the Atomic Energy Commission a reexamination of the question of the genetic effects of the atomic bombs, this time employing biochemical phenotypes as indicators of mutation. The systems were to be the same as those being employed so extensively in the Indian studies. This proposal was approved by the American side, but its full implementation has been blocked by the Japanese side until the present time. Nevertheless, a small Pilot Study has been initiated, which has entailed organization of a screening laboratory, the training of 3 Japanese professional and 6 technical personnel,

and examination each week of some 48 blood specimens reaching the ABCC in the course of approved protocols. In connection with this Pilot Study Tanis, Ferrell and Neel have each spent 4-6 weeks in Japan during the past year. The crux of the proposed study is a search, in children of A-Bomb survivors and of unirradiated controls, for variants not present in either parent and so presumably the result either of mutation or non-paternity. The latter can be detected with high accuracy by appropriate typings.

However, the control material for this study will also provide the basis for an indirect estimate of spontaneous mutation rates in the Japanese, similar to that made for the American Indian. Earlier we referred to the unexpectedly high estimate of mutation rates in the Amerindian. There are many difficult statistical problems in deriving a comparable estimate for Caucasian, Negro or other Mongoloid groups, involving especially the estimation of N and $\bar{t}_0$, and it is clear there will be some lively debates in the years ahead. There are also serious biochemical problems in comparing the output of different laboratories. We propose to proceed cautiously, but find the data already at hand quite thought-provoking. Table 5 summarizes the results to date from the ABCC Pilot Study, together with our own data on the Indian populations and what appear to be the most nearly comparable data for Caucasians. For the latter, the enzyme data are those of Harris and colleagues (1974), with various other groups (or summaries) providing the information on the three serum proteins and hemoglobin.

There are some very real difficulties in comparing the data on Indians with the other two sets. The method of sampling Indians produces "clusters" of variants, but when one encounters as large a "cluster" as is created by the Yanomama-2 albumin variant, it completely unbalances the sample. For now we tentatively conclude that the frequency of rare variants is about the same in these three ethnic groups.

If this conclusion is sustained, it has some interesting implications for mutations rates in the three groups. Both the size of the breeding population (N) and average mutant survival time ($\bar{t}_0$), which appear in the denominator of the equation used for calculating mutation rates, should be larger for Caucasian and Japanese populations than for Indians. This implies a lower mutation rate in these two groups.

Table 5. A comparison of the frequency of "private" variants of 15 proteins in three ethnic groups.

Population	Alb	Tf	Crpl	Hgb	6-PGD	PGM ₁	PGM ₂	LI-H	AK	MDH	ADA	Pep A	Pep B	pH1	ICD	Av ⁽³⁾ Enz	Av Prot	Av All
Japanese	$\frac{5}{2251}$	$\frac{49^{(1)}}{2250}$	$\frac{1}{2251}$	$\frac{1}{2251}$	$\frac{1}{2233}$	$\frac{5}{2246}$	$\frac{0}{2247}$	$\frac{3}{2250}$		$\frac{3}{2250}$	$\frac{0}{2232}$	$\frac{1}{2227}$	$\frac{3}{2251}$	$\frac{16}{2249}$	$\frac{1}{2218}$			
Per 1000	2.2	21.0	0.4	0.4	0.4	2.2	0	1.3		1.3	0	0.4	1.3	7.1	0.4	1.4	6.0	2.8
Indian	$\frac{60+491^{(2)}}{6000}$	$\frac{27^{(1)}}{5559}$	$\frac{20}{4604}$	$\frac{0}{3989}$	$\frac{0}{4769}$	$\frac{0}{4725}$	$\frac{0}{4731}$	$\frac{0}{3070}$	$\frac{0}{3936}$	$\frac{0}{2895}$	$\frac{0}{2541}$	$\frac{13}{2882}$	$\frac{0}{2944}$	$\frac{4}{2937}$	$\frac{8}{2917}$			
Per 1000	10.0- 92.0	5.0	4.3	0	0	0	0	0	0	0	0	4.5	0	1.3	2.7	.9	4.8 -25.0	2.0 -7.8
Caucasian ⁽⁴⁾	$\frac{8}{9890}$	$\frac{5}{1007}$	$\frac{12}{624}$	$\frac{10}{8000}$	$\frac{1}{4939}$	$\frac{12}{10,333}$	$\frac{7}{10,333}$	$\frac{2}{1015}$	$\frac{1}{6760}$	$\frac{1}{516}$	$\frac{2}{4798}$	$\frac{8}{8798}$	$\frac{15}{7041}$	$\frac{1}{1553}$	$\frac{4}{718}$			
Per 1000	.8	5.0	19.2	1.2 (Sick et al)	0.2	1.2	0.7	2.0	0.2	1.9	0.4	0.9	2.1	.7	5.6	1.6	6.5	2.9

Notes

(1) Japanese data include 29 D_{Chi} (Japan) and Indian data, 27 D_{Chi} (Piaroa)

(2) With and without Yan-2

(3) No AK

(4) Caucasian refs. - from Harris et al. except Hgb = Sick et al (1967)

Alb = Fine (1970)

Transferrin = Rex-Kiss and Fésűs, L. (1970)

Ceruloplasmin = Bajatzadeh and Walter (1969)

If the spontaneous mutation rate in Indians really is higher, it could be because of genetic (ethnic) or environmental factors. The preliminary data from Japan provide a partial control on ethnicity, since both Indians and Japanese are (long-separated) Mongoloids. Thus at the moment, we are inclined to seek an extrinsic factor for the possibly higher mutation rate of Indians, perhaps related to the factor causing the chromosomal damage described in earlier reports.

In our opinion, if the thesis under development proves correct, it has far-reaching theoretical and practical implications. From the practical standpoint, geneticists have in recent years repeatedly expressed concern over the possibility of mutation rates increasing from exposure to radiation or mutagenic chemicals. These are, of course, legitimate concerns, but if they are superimposed on a background of generally decreasing mutation rates (? because of the control of many viral diseases, improved food preservation, etc.), then the problem of arriving at a balance between gain to society from the processes resulting in these exposures and harm to the individual has assumed a new dimension. On a more theoretical level, if this thesis is correct, it would indicate that in a higher organism the basic process of mutation may indeed be substantially altered by exogenous influences, and should spur a search for the nature of these influences.

V.2.2. Laboratory Developments Related to Population Monitoring (Tanis, Ferrell, Neel). - During the past year, a portion of the effort of Tanis and Ferrell has been devoted to adapting additional starch gel systems for use in screening for rare variants. As of this writing, the following six additional erythrocyte protein systems appear suitable:

- a) 2,3 diphosphoglycerate mutase. Method of Chen, S. H., Anderson, S. E. and Giblett, E. R. (1971).
- b) Aldolase. Method of Charlesworth, D. (1972).
- c) NADH diaphorase. Method of Hopkinson, D. A., Corney, G., Cook, P. J. L., Robson, E. B. and Harris, H. (1970).
- d) Nucleoside phosphorylase. Method of Edwards, Y. H., Hopkinson, D. A. and Harris, H. (1971).
- e) Galactose-1-phosphate uridyl transferase. Method of Weitkamp, L. R. (1973).

- f) Triose phosphate isomerase. Method of Peters, J.,
Hopkinson, D. A. and Harris, H. (1970).

Because of different buffer and gel potential requirements only two systems can be combined without compromising band pattern quality; the two are NADH diaphorase and galactose-1-phosphatase uridyl transferase.

Also during the past year, Tanis and Ferrell have been involved in a collaborative effort with Dr. Charles Scott and Dr. Tom Tiffany of the Oak Ridge National Laboratory regarding steps to automate aspects of the laboratory procedures involved in electrophoretic screening. The first step being automated is the process of sample preparation. A prototype centrifugal cell washer has been tested which will yield in continuous process plasma samples, washed red cells for long term storage, and hemolysate for immediate laboratory use. The prototype has been tested for one month under laboratory conditions, and the ORNL group are now undertaking modifications which should lead to a model suitable for routine laboratory use.

In addition to the cell washer development, there have been discussions concerning the possibility of automating the gel casting process. This would probably involve a conversion of the laboratory to the exclusive use of polyacrylamide gel electrophoresis. The recent introduction of a commercial polyacrylamide gel electrophoresis system based on the use of pre-cast, standardized gels is of interest in this respect.

V.3. MATHEMATICAL AND STATISTICAL ASPECTS OF POPULATION GENETICS

V.3.1. Model Development.

V.3.1.1. General Introduction. - The proper analysis of genetic data on human populations involves the selection of the appropriate model, relating the genetic parameters of interest to the observations which can be made, followed by the selection of the appropriate statistical functions of the data to estimate parameters of the model. We have been concerned with both of these considerations in our efforts to identify the causations of the allelic variation occurring at one-third to one-half of

the gene loci in human populations. The causations of the observed dispersion of allele frequencies among subpopulations of man may be categorized into the demographic and the functional. Migration, sub-population size, mating preference and age-distribution are all factors which may have a demographic influence on allele frequencies. Mutation, linkage and the effect of allelic differences or the effects of interactions between alleles at different loci on the survival and fitness of a phenotype are all factors which play a functional role in determining allele frequencies in these subpopulations. During the past year we have continued our efforts to evaluate genetic models which relate these two classes of factors to the variation of allele frequencies, and to develop statistical procedures which estimate the genetic parameters of these models. New models have been applied to data arising from our studies of human populations and to data simulated by the computer. The latter is a most important step in evaluating the properties of the statistical procedure of interest in a situation where the genetic parameters can be specified.

V.3.1.2. A Statistical Model for the Analysis of Genetic Parameters Which Define Human Population Structure (Rothman, Sing, Templeton).

- The variation of allele frequencies among subdivisions of a population (such as the Yanomama) may be considered a function of mutation, migration, sub-population size and selection. Three genetic parameters may be estimated for each locus from genotype frequencies which reflect the operation of these causative factors. They are the set of allele frequencies, the correlation between alleles within individuals (the inbreeding coefficient) and the correlation between alleles in different individuals but in the same subdivision of the population. Knowledge of the values that these parameters take in a human population yields a description of the consequences of the operation of these various causations on the dispersion of allele frequencies among subpopulations. Estimates of these parameters in a natural population provide the basis for prediction of the evolutionary future of allelic variation in replications of the populations and a scheme for testing hypotheses about the forces which influence such changes. We have prepared a paper (Appendix 12) which presents a

theory for estimation of these three genetic parameters, utilizing the maximum likelihood principle. Although our model is not intended to evaluate the relative roles of the alternative forces which may lead to the structure of a population, it does provide the procedure to separate those loci with allelic variations that result from mutation, migration and population size (causations deemed most relevant by the neutralist school) from those loci which have the additional influence of selection determining their allele frequencies. Briefly, our likelihood model is based on the utilization of a multivariate negative binomial distribution for the genotype frequencies within and among subdivisions of a population. This distribution is a function of allele frequencies and the two kinds of correlations; correlations among alleles within individuals and correlations between alleles in different individuals of the same subdivision. This is a fundamental model which can provide the basis for developing a spectrum of statistical models which can incorporate the role of a number of relevant genetic parameters. In Section VI.4.1. of the Proposal we present our plans for the extension of this model to multiple loci, so as to distinguish loci which are under the influence of selective pressures from those which are not.

V.3.1.3. Demographic Models Which Incorporate Density-Dependent Vital Rates (Smouse, Weiss). - Efforts to develop a model for density-dependent demography culminated in the manuscript attached as Appendix 13. The basic finding so far has been that the stable age-distribution is very much dependent upon the type of density-dependent feedback employed. We investigated three different feedback strategies: 1) density-dependent fertility, 2) density-dependent survival, and 3) density-dependent post-infant survival. We found increasing steepness of the age-distribution as one moved from strategy (1) to strategy (3). The details are described in the manuscript.

In the study above, we investigated a model with high innate capacity for increase. The growth trajectory of the population, in the absence of density-dependent feedback, is described by:

$$N(t + 1) = \lambda R(t),$$

where $0 < \lambda < 2$. The model employed has $\lambda = 1.946$, a near maximum. Recently, we have investigated the impact of intrinsic growth rate (λ) on the age-distribution. To do this, we have investigated the same three density-dependent feedback strategies for similar models with growth rates of $\lambda = 1.002$ (barely sufficient for growth) and $\lambda = 1.514$ (medium growth). The steady-state age distributions are shown in Table 6. The potential array of stable-age-distributions is much more limited in populations of low intrinsic growth rate than in populations of high growth rate.

We have also investigated the rate of convergence to equilibrium for these models and for a variety of others. In particular, we have exhaustively investigated two different types of periodic behavior, arising from periodic fertility damping. We have related such models to comparable periodic cases known from classical density-independent demography.

V.3.1.4. A Comparison of the Results of Using Diffusion Equations and Branching Chain Processes In Genetic Problems (Jensen). - In several recent papers, Kimura and Ohta (1969 a,b) and Maruyama (1972) have studied theoretically, and with simulations, the problem of computing the average number of generations until fixation or loss of a new mutant in a finite population. The theoretical model used was the diffusion equation approximation to the classical Fisher-Wright model for the distribution of gene frequencies in finite populations with non-overlapping generations.

The same problem is being studied for finite populations by using branching processes. This is similar to the technique used by Fisher (1930), except that he was concerned with infinite populations. A branching process model can be used to describe mathematically the distribution of gene frequencies in a constant size adult population by appropriate methods of conditioning. The basic assumption is that the adults produce offspring according to a known distribution; mutations, or new genotypes are then introduced into the offspring of the parents. A sample, with or without replacement of the total number of individuals present (parents plus offspring), is then taken to produce the next generation. The process is then repeated.

With very simple assumptions on this type of reproduction, it can be shown that the lifetime of a surviving (mutant) offspring is distributed

Table 6: Stable age-distributions for a five-class population under three different density-dependent damping regimes and three different intrinsic growth rates.

Growth Rate (λ)	Damping Strategy	Stable Age-Distribution				
		F ₁	F ₂	F ₃	F ₄	F ₅
1.946	I*	.2040	.2020	.2000	.1980	.1960
	II	.5087	.2587	.1316	.0669	.0341
	III	.9900	.0099	~ .0001	~ 0	~ 0
1.514	I	.3278	.2458	.1844	.1383	.1037
	II	.5087	.2587	.1316	.0669	.0341
	III	.7500	.1880	.0472	.0118	.0030
1.002	I	.5072	.2589	.1321	.0674	.0344
	II	.5087	.2587	.1316	.0669	.0341
	III	.5103	.2586	.1310	.0664	.0336

*I = Density-dependent fertility

II = Density-dependent survival

III = Density-dependent post-infant survival

geometrically. If every parent is credited with one-half of each offspring, the mean lifetime is two units of time with variance also two units. This differs with what is reported in the literature. Maruyama (1972), using the diffusion approximation, shows that the mean lifetime is two units of time with a variance of four units. There are problems to be reconciled.

At present, different probability models for the distribution of offspring and mutations are still being investigated. The resulting mathematical models have a very complicated structure and various approximations are being sought to make interpretations easy. It also remains to compare these new theoretical results with observed results on the survival of a mutant.

V.3.1.5. Evolution in a Heterogeneous Environment (Templeton, Rothman). - Levins introduced the concept of fitness set and adaptive function to analyze evolution in heterogeneous environments. His specification of a adaptive function has been shown to be a special case of Bayes rules, and generalizations of these rules in "fine" and "coarse" grained environments have been investigated in the framework of a no-data decision theory problem (Templeton and Rothman, in press). Particular attention is paid to the concept of maximum strategies (i.e., the minimum fitness over environments is maximized).

V.3.2. Evaluation of Models and Estimation of Parameters.

V.3.2.1. Evaluation of the Likelihood Statistical Model For Analysis of Parameters Which Measure Population Genetic Structure (Rothman, Sing). - A comparison of the likelihood approach developed above (Section V.3.1.2.) with the least squares estimation approach developed by Cockerham (1969) was made using two sets of data. First, the six largest kindreds of the Tecumseh Community Health Study, defined by Sing et al (1973), were chosen to represent a random sample of subdivisions. Data from five co-dominant loci, each with two alleles, were available from individuals from each kindred. Second, a computer simulation was employed to generate 100 replicates of six independent samples of data from a reference population defined by specific correlation parameters, subdivision size and genotype frequencies. The six subdivisions (of equal size) were then subjected to

the two analyses. On the average, the variance of estimate of the correlation between genotypes within subdivisions (ρ), using the least squares procedure, was double that obtained using maximum likelihood procedure.

For the Tecumseh data we found that the least squares estimate of ρ based on the Ss and the haptoglobin markers appear to be substantially different from zero, whereas three other marker loci gave estimates which were not significantly different from zero. The results from the application of maximum likelihood and least squares estimation procedures to the output of the computer simulations are given in Table 7. Increasing subdivision size reduces the percent bias of the maximum likelihood estimate. However, regardless of the estimation procedure, for the smaller values of true ρ , the biases are affected little by increases in subdivision size from 20 to 100. For values of ρ expected in real sexually reproducing populations (say $\rho \leq 0.1$), it appears that each of the estimation procedures considered will err badly for small subdivision sizes.

The dispersion of estimates generated by simulation yields a somewhat different comparison. The coefficient of variation reflects a decrease in the relative variance for all subdivision sizes as true ρ increases. Regardless of the subdivision size or value of ρ considered, the variances of least squares estimates are more than double the corresponding variances of the maximum likelihood estimates. The combined influences of bias and variance are reflected in Table 7 by the mean squared loss measure. By this criterion the maximum likelihood and ρ_0 estimators are equally good and two to twenty times better than the least squares estimator.

V.3.2.2. Estimation of Factors Contributing to Gametic Disequilibrium in the Hutterites (Martin, Sing). - One strategy for determining the role of functional (non-random) effects on the organization of the gene pool is to study the correlation between non-alleles in gametes. It is widely accepted that the phenotype is the unit of selection. However, little research in human populations has been conducted to estimate the size of the genetic unit of selection. Sinnock and Sing (1972a, b) found that 27% of the gametic disequilibrium values estimated using 11

Table 7. Summary statistics on $\hat{\rho}$ computed according to three procedures for four combinations of parameters each replicated 100 times.

Estimator	N_h	Mean			% Bias ⁺			Variance x 10 ⁺³			Mean squared loss		
		20	50	100	20	50	100	20	50	100	20	50	100
True $\rho = .1$													
ρ_0		.081	.082	.088	19.2	18.2	11.9	2.98	1.67	1.68	0.0033	0.0026	0.0003
ρ_{MLE}		.084	.085	.086	16.2	15.4	14.1	3.15	1.86	1.20	0.0034	0.0021	0.0002
ρ_{LS}		.108	.093	.108	8.2	7.4	7.9	7.41	3.83	3.23	0.0078	0.0038	0.0021
True $\rho = .25$													
ρ_0		.193	.217	.214	22.8	13.2	14.5	7.95	8.77	6.01	0.0112	0.0099	0.0073
ρ_{MLE}		.196	.217	.212	21.8	13.3	15.1	6.27	6.96	5.10	0.0092	0.0081	0.0065
ρ_{LS}		.221	.234	.256	11.4	6.4	2.3	16.66	14.94	12.70	0.0175	0.0157	0.0127

$$+ \frac{\hat{\rho} - \text{True } \rho}{\text{True } \rho}$$

polymorphic loci were significant at the 5% level of probability. Selection and subdivision effects were cited as possible causes. Since this first work (in any human population) we have compared our result with gametic disequilibrium values computed from a set of data collected on a California population (Sing et al., in press). No consistent deviations from $D = 0$ for the two studies were noted. We (Sinnock and Sing, 1972a, b) pointed out the possible effects of subdivision in Tecumseh but could find no logical procedure for separating such effects from the residual functional influence of selection. The Hutterites provide a sample which gives us the opportunity to remove the effects of subdivision and small population size (demographic influences) from gametic disequilibrium. Table 8 lists the values of D' (defined in Table 8) for the pooled analysis of the four clans which make up the S-Leut. The question of interest is: "To what degree are the large and statistically significant values of D' due to the subdivision of the Leut into clans?" Because the ancestry of each individual is known, it is possible to estimate the role of such demographic influences. We plan to continue this study in the forthcoming grant period (see Section VI.4.4.).

V.3.2.3. Analysis of the Influence of Genetic Polymorphism at a Single Locus on the Distribution of Cholesterol Values (Sing). - During the past year we have continued to evaluate the relative role that genes may play in determining the normal cholesterol variation within a total community. We have available 14,310 observations from the Tecumseh Community Health Study. This body of data has provided the basis for our continued interest in the relative role that genetic factors play in determining quantitative variation in those variables which may be used as predictors of coronary heart disease. The current favored hypothesis (Goldstein et al. 1972) is that individuals who are hypercholesterolemic (exceed the 95th percentile) are either homozygous or heterozygous for a dominant allele at a single genetic locus. Work completed during the past year (Sing et al., in press) was aimed at first adjusting the observations for age and temporal effects and then using the adjusted values in an analysis to determine the contribution of a single polymorphic locus. Approximately 40% of the total variation could be removed by adjustment for age and temporal effects. The analysis of the adjusted data indicated

Tab . Estimates of D' for the S-Leut of the Hutterites ($D' = D/\text{max possible } D$ where $D = \text{covariance between nonalleles}$).

	MALES: Parents below diagonal and offspring above diagonal						
	A	B	M	K	Hp ¹	Gm ^a	Gm ^b
ABO : A			-.1809**	.0749	.0716	.0150	.0002
B			-.3773**	-.1359	.3536	.0513	-.0386
MN : M	-.1281**	-.5664**		.0258	-.0170	-.0096	.0166
KELL : K	.3164**	.1539**	-.0184		.1188*	.0125	-.0573
HAPTOGLOBIN: Hp ¹	.1070	.3905	-.0068	.0661		-.0234	.0381
GM : Gm ^a	-.2647	-.2042	-.0821	-.0092	-.1691		
Gm ^b	.1759	-.0852	.1047	-.1002	.1567		
	FEMALES: Parents below diagonal and offspring above diagonal						
	A	B	M	K	Hp ¹	Gm ^a	Gm ^b
ABO : A			-.1080*	.1041	.0691**	-.0183	-.0305
B			-.0819*	-.1965	.6733**	-.7445	.5945
MN : M	.0780*	-.7252*		.0547	-.0769	.2223*	-.1342*
KELL : K	.1864*	.9583*	-.0789		.0087	.0690	-.0687
HAPTOGLOBIN: Hp ¹	-.0214	.3307	.0182	.0260		.0291	.0051
GM : Gm ^a	-.0909*	-.8846*	-.0355	.1064	-.0261		
Gm ^b	-.0560*	.9043*	.0171	-.0429	.0763		

*Significant at the .05 level of probability; **Significant at the .01 level of probability.

that there is little evidence that the frequency of the hypothetical cholesterol-elevating allele in the Tecumseh population is as high as estimated by Goldstein et al. (1972). We found that only 1 of the 42 families ascertained by an individual exceeding two standard deviations of the population mean had both an affected parent and an affected child. Table 9 presents the recurrence of hypercholesterolemic levels in the families which had an index case. The low frequency (1.5%) of affected siblings in families ascertained through the parents does not support the segregation of a dominant gene as a major source of high values. On the other hand, there is evidence for the role of a familial component determining the normal variation in cholesterol among siblings. Table 10 gives an analysis of the quantitative variation among parents and children for the 1228 nuclear families with both parents and at least one child measured for serum cholesterol. Overall, the correlation between mid-parent and oldest child (0.35) is significantly greater than between parents (0.12). The significant intra-class correlation (0.42) among sibs is also evidence for the role of a familial component.

V.3.2.4. Estimation of the Contribution to Blood Pressure Variation of Socio-Demographic Changes in a Cohort From the Tecumseh Community and the Relationship to Changes in the Genotype Distributions (Sing). - Along with cholesterol, blood pressure is one of the major biological measurements which can be used to predict coronary heart disease. D. Smith (post-doctoral fellow) is working with Sing on an analysis of factors contributing to the variation of blood pressure in a total community, Tecumseh. The sample of blood pressure values taken over a period of 12 years from the Tecumseh community provides an opportunity to study a cohort of individuals over time. One may determine the effects of the socio-demographic changes on the frequency distribution of blood pressure values. Any changes in the frequency distribution which take place in the cohort can be related to variation at 12 genetic loci in addition to a large number of socio-economic variables. The frequency distribution of blood pressure in the 1960 cohort had changed significantly by the time the cohort was reexamined, three years later. It changed from a distribution which was positively skewed to one which was more normally distributed. The loss

Table 9. Frequency analysis of 73 nuclear families having one or more members with an adjusted cholesterol level 2 S.D. above the mean.

Family ascertained by	No. of families	No. of parents > 2 S.D.	No. of children	No. of children > 2 S.D.
Parents	32	33	66	1
Children (all)	42	1	110	46
Simplex cases	38	1	96	38
Multiplex cases	4	0	14	8

Table 10. Summary statistics for the 1228 nuclear families with both parents and at least one child measured for serum cholesterol.¹

Group	N	Parents						Children							
		Father			Mother			r^2	N				$\hat{\sigma}_{W^3}^2$	R^4	r^5
		$\bar{x}$	$\pm$	S.D.	$\bar{x}$	$\pm$	S.D.								
		$\bar{x}$	$\pm$	S.D.	$\bar{x}$	$\pm$	S.D.			$\bar{x}$	$\pm$	S.D.			
All families	1228	5.30724	.1718	5.29796	.1649	.12	2767	5.29964	.1634	.0170	.42**	.35*			
Child index	42	5.37358	.1319	5.34542	.1540	-.20	110	5.19394	.2054	.0476	-.24	.07			
Parent index	32	5.49778	.2164	5.57065	.2249	-.69**	66	5.33615	.1754	.0229	.28	-.53*			

¹Adjusted for cholesterol (Base 6).

²Product moment correlation between parents.

³Estimate, from one way analysis of variance, of variance among sibs in families.

⁴Intra-class correlation; among sibs in families.

⁵Product moment correlation between average of parents and eldest sib.

*Significant at the .05 level of probability.

**Significant at the .01 level of probability.

from the cohort of individuals with the higher blood pressure values was due, in part, to movement away from the community and, to a lesser degree, to mortality. The frequencies of marker phenotypes of those who stayed and those who left was not significantly different for any of the 11 genetic systems studied. We are continuing to analyze these data to determine which socio-economic variables are correlated with the higher blood pressure values of those who left between the third year and sixth year measurements.

V.3.2.5. Statistical Analysis of Genotypic Structure of Natural Populations (Smouse). - As part of a long-range project, we have continued to develop statistical techniques particularly suited to the analysis of genotypic structure in natural populations. Our efforts for the past year have been on two fronts.

1. The analysis of clinal variation in allelic frequencies. Smouse and Kojima (1972) described multinomially based, likelihood estimation and testing procedures for clinal variation in allelic frequencies. The model was restricted to linear trends and single loci. We have now extended this analysis to sigmoidal trends and multiple loci. The analysis has been applied to several different examples, and is quite useful. The resulting paper, which will appear in Theoretical and Applied Genetics, is included as Appendix 14. The techniques developed in this paper have also been used by the Yanomama group to examine gene frequencies with some success (see Appendix 11).

2. The analysis of recombinational disequilibrium in multiple-locus gametic frequencies. We have developed a multinomially based, likelihood ratio analysis of multiple-locus gametic disequilibrium. This analysis is based upon recent developments in contingency analysis, which are particularly suited for the description of gametic disequilibrium. The resulting paper will appear in Genetics, and is included as Appendix 15. Two findings were of primary interest, and are worth mentioning briefly. It was discovered that neither the estimates nor the χ^2 test criteria for all two-locus comparisons are independent. Also, disequilibrium between unlinked loci does occur, although that between linked loci is quite a bit larger.

V.4. POPULATION SIMULATION (Li)

As reported in past applications and publications, we have a heavy investment in computer simulation of tribal-type populations. We are operating on the premise that a number of significant questions concerning the genetic consequences of the population structure we are studying are best approached through Monte-Carlo simulation. For instance, we have determined average survival time for a mutant gene; the usefulness of estimates of this parameter in the indirect estimate of mutation rates has already been brought out (Section V.1.4.).

The simulation model in use up to the present time was developed by MacCluer, Neel, and Chagnon (1971). Imperfections in this first model were that the complex of 4 villages being simulated constituted a closed community (no migration in or out) and there was no provision for village fission. A consequence of the latter, given the system of marriage alliances written into the program, was that most of the population ultimately became concentrated into that village which was largest at the outset. Finally, the demographic data available at the time the original simulation was constructed were less precise than those currently available (Section V.1.8.).

Li has devoted much of the past year to improving the simulation. Provision has been made for immigration and emigration, for village fissioning when a village size of 300 is exceeded, and the life tables which are incorporated into the new model have been modified to correspond with the recent analysis of Neel and Weiss (Section V.1.8.). Furthermore, since a 400-year run was planned, and since the program called for a 0.6% increase per year in population (the present Yanomama rate), there were special problems in programming so as not to exceed the core or storage capacities of our present computer.

Because it was anticipated that each 400-year run would be quite time-consuming, it was clear that a number of features should be built into each run. The three of primary interest are:

1. Inbreeding as measured by identity of descent. Each individual in the input population has been assigned a pair of unique alleles at each of four loci. At the termination of the run, if the alleles at some locus in an individual are the same, it can only be because of identity by descent.

2. Founder effect. The design referred to in the preceding paragraph also permits an exhaustive analysis of the founder effect. Specifically, the number of alleles surviving at the end of a run and the frequency of each will permit an evaluation of the cumulative effects of Yanomama social structure on mutant survival.

3. Zygotic disequilibrium. One of the current searches for evidences of genetic selection requires an examination of whether the genotypes at one locus occur in random combination with the genotypes at a second locus. Significant departures from randomness are termed zygotic disequilibrium and may, when certain prerequisites are met, be taken as evidence of the action of selection. It is proposed to examine the Yanomama data for such disequilibrium. However, it is anticipated that the population structure of the Yanomama may of itself result in considerable disequilibrium. The structure of the simulation (which does not include selection) will enable us to construct a null expectation for disequilibrium, with which to contrast the actual data.

At this writing Li has just completed the first production run. The model appears to be satisfactorily reproducing the demographic characteristics of the Yanomama. It is anticipated that within the next several months, a number of the above analyses will have been carried out on the results of this and additional runs.

V.5. UPGRADING THE COMPUTER

V.5.1. Hardware. - At present, the analyses being supported by the AEC Program in Population Genetics require more computing time than is available on the IBM 1130 system purchased for this Program by the AEC. This IBM 1130 computing system is a disk-oriented mini-computer which consists of a central processing unit, a card reader-punch, a printer and a Calcomp DS12 disk drive. During the past year we have conducted an extensive evaluation of our present computing situation and our needs for the next five years, and have surveyed the hardware and software necessary to meet our projected needs. A report (Appendix 16) was submitted to AEC in January, 1974. As our program in population genetics evolves, our computing needs not only increase, but change in character. The two supplementary proposals attached to this application reflect some of the

new demands on our computing facility. As an example of new directions, we anticipate that rather than "batching" our data from the lab bench, through the analyses to the final report, we will feed our data directly from the lab bench into a storage unit. Rather than punching cards and building a deck which represents the information available from a study, each observation will be fed directly from the measurement instrument or a teletype terminal. The investigator will then utilize the full power of the computer to update, verify, and correct his data, and conduct ongoing analyses. This will enable him to make judgments during the progress of the study which would not otherwise be possible until the data had been gathered together and batched into a storage unit. Our proposal, given in Appendix 16, reflects an increased capacity both in computing and the ability to conduct realtime data acquisition, storage, retrieval and analyses, which cannot be handled by our IBM 1130.

V.5.2. Software. - We have been developing a set of programs to handle laboratory and clinical data for the Sickle Cell Center with which the Department of Human Genetics is involved. Currently, this operation is being handled by the University's IBM 360. Any new projects for which we collect data, such as described in Supplementary Proposals 1 and 2, will require a similar system of programs to handle the dynamic acquisition and analysis of the data. Thus, the experience with the Sickle Cell Project should be quite useful if we move to implement Supplementary Proposals 1 and 2.