INTRODUCTION

Analysis of surnames for genetic purposes has been suggested (1) and widely used (2) as a shortcut for estimating the frequency of cousins, and hence for evaluating the inbreeding coefficients of populations. Surname frequencies have been employed, in lieu of gene frequencies, as a source of information for the study of isolation by distance (3). Our interest here is slightly different, motivated in part by a series of earlier investigations (4,5,6,7,8,9,10) in which the distribution of surnames in a population was equated to that of alleles at a locus. In this respect, surnames offer a very rich source of information that can be tapped to answer specific problems of genetic population structure.

In the great majority of populations surnames are transmitted via the male line. Transmission is almost like that of Y-chromosome genes except that surnames are also passed to females, who usually maintain that surname at least until they are married. A notable exception in the Western world is the Icelandic population, which maintains the custom formerly prevalent not only among Scandinavian but also among Celtic populations: the surname is derived from the first name of the father and therefore changes at every generation. In societies where clan names are used, the names are usually transmitted patrilinearly and are therefore essentially equivalent to surnames.

There is considerable global variation in the time of first use of surnames. In most of Europe the use of modern surnames began during the late Middle Ages and spread to the whole population by the beginning of the Renaissance. In Japan they were not used until the last century. In China they are probably over 4000 years old, as will be indicated in more detail in section 2. This is the earliest known use of surnames.

In this paper we analyze surnames obtained from the stratified random sample covering a 1/2000 fraction of the whole Chinese population as censused by the Government of the People's Republic of China in 1982. The

sampling units were entire "administrative villages" or parts of cities ("residents' committees"). We pay special attention to the distribution of surnames thus sampled, its interpretation as that of selectively neutral alleles at a single haploid locus under uniparental transmission, and the use of the parameter estimates obtained from this distribution for the analysis of genetic population structure. We compare the distribution with that obtained from genetic data.

A related paper, "Congruence between Genetic and Linguistic Evolution in China," has appeared in the Journal of Chinese Linguistics. The article addresses implications of the analysis of Chinese surnames for the study of Chinese linguistics. The main conclusion of the comparison with linguistics is that there is a highly significant statistical correlation between linguistic and surname distances. This correlation probably arose because both linguistic and surname distances reflect migration patterns. Although too few genetic data were available for comparison with linguistic data, they are expected to correlate highly with surname data due to parallels in transmission of genes and surnames. A second major conclusion is that Chinese surnames and language data are more correlated with each other than they are with either geographical or historical information.

1. METHODOLOGY: THE USE OF SURNAMES FOR THE STUDY OF POPULATION STRUCTURE

a) Distributions of the Numbers of Surnames.

Surnames can be considered as alleles at a single locus, and allele distributions can be applied to distributions of surnames. Allele distributions are difficult to study in the case of genetic alleles because the sample of alleles of a gene that are detectably different by usual techniques is extremely small (rarely more than five or six per locus). Also, the number of