History of Human Migrations From Genetic Data

Prasanth Sankar

1 Introduction

Origin of human beings and their migrations before the dawn of historical records has fascinated anthropologists and evolutionary biologists for years. Till the immediate past most of the studies used fossil records. The emphasis is on the morphological continuity of fossil records and their geographical correlations[?]. Fossil data records and their analysis seems to indicate a multiple origins of human beings in each continent. But recent reexamination of the fossil data contests this hypothesis but is unable to provide any alternative pictures that can be corroborated. Another traditional approach has been to use language and linguistics to recognize possible prehistoric connection between people and to suggest certain past migratory events. The limitations with this method are, to get detailed information about the historic events presence of written or oral records are necessary and environmental influence on the development of present day languages is not at all recognized. This leaves scientists to look for better alternatives as indicators of historic events which are accessible presently and which can be subjected to experimental unraveling.

With the advent of genome sequencing and statistics to interpret and infer the information contained in them human population genetics has started using distribution of genetic markers in extant human populations to gain insight into demographic and migrational history[?]. Here the principal aim is to identify those regions in genome that faithfully reproduce information regarding genetic drift and migration and is less affected by natural selection brought about by random mutations. For this, those regions should be passed from parent to offspring with out the shuffling effect of recombination and should have less recurrent mutation. The mutation effects should be of such nature that the regions of varying ages should be related by a clear step wise mutation process that is geographically correlated. Here it should be noted that the choice between selection and major migration routes can be made only on the basis of possible environmental correlations that would favor selection or of clearly repeated patterns seen for several unrelated genes that would support migratory explanations[?]. Two such genetic regions used for a long time for this purpose are, Mitochondrial DNA and HLA variation(HLA region contains a number of closely linked highly polymorphic genes whose products control a variety of functions concerned with the regulation of immune response. The gene distribution records the evolutionary history of disease infection). These genetic studies point to human origin in Africa, with subsequent dispersal of modern humans throughout the rest of the world. Even though this is accepted without much criticism the number of migrations out of Africa is still unclear and further studies are needed to identify the number of migrations and the geographic route taken by each. In these studies, to infer about the time periods of an event one has to make certain assumptions about the mutation rate and mutation processes (single mutation or multiple mutations etc.). In most cases these events are modeled on a computer and time elapsed since a major mutation associated with a particular past event is infered.

In this paper we review four of the recently introduced methods to study this dispersal of human population from Africa. The first method uses the diversity of the non recombining portion of the Y chromosome(NRY)[?]. Like mtDNA this region of Y chromosome is passed from parent to offspring (father to son in this case) without the shuffling effect of recombination which allows the evolution and retention of a wide variety of stable genetic determinants(haplotypes) with varying ages, through a clear, stepwise mutation process. It is also noted that NRY diversity within populations is lower than that seen for other markers. These make NRY a powerful tool for historical and demographic studies. NRY data is used to shed some light on the path followed by humans after they

left Africa and settled other continents. Here evolutionary relationship among NRY haplotypes is used to infer details of past migrations. The second method[?] uses haplotype systems that contain preferentially densely spaced single-nucleotide polymorphism(SNP),in this case a 565 bp region of chromosome 21 ,which have low mutation rates and essentially no recombination. The third method explicitly uses the results from Human genome maps and progress in that field and is conforming to the present spirit of application of computers to the analysis of genome data(Bio informatics)[?]. This method uses genome-wide map of single nucleotide polymorphism(SNP). Correlations among the neighboring alleles, known as linkage disequilibrium(LD) which reflect haplotypes descended from a single, ancestral chromosome is used to study the evolutionary history of humans. The last method uses the study of evolutionary history of a certain virus. It is assumed that humans in their migration carried this virus along with them and a comparison of certain genetic markers of the virus from different geographic regions can give insight into human migrations. One such study uses hepatitis G virus [?].

2 Methods

In the case of NRY study blood samples were collected from people from different geographic regions such as Africa, Asia, Europe and America. DNA is purified from these samples and Polymerase chain reaction(PCR) was used to pinpoint the selected allele. Ancestral and derived states were determined by comparisons to nonhuman primate sequences. Haplotype frequencies and diversities were determined from the PCR data. The age of the haplotype is estimated by using a single step mutation model and the haplotype frequency variation was used to construct a population tree. In the case of Chromosome 21 region a variation of this method is used.

In the linkage disequilibrium study the SNP's were identified by bio informatics and appropriate statistical techniques were used to determine the correlations(LD) between them. The analysis were carried out on DNA data from North Europeans and Nigerians. A computer simulation was also carried out assuming that the population experienced an extreme founder effect or bottleneck taken into account by increased inbreeding and the LD were calculated in this case too.

In the viral case genome sequences of the virus from various geographical regions were subjected to bioinfomatic analysis to determine the presence of certain genes whose presence and distribution is found to be correlated to the geographic region. The selected region of DNA was subjected to sequence comparison and Phylogenetic tree of the virus was constructed.

3 Results and Discussion

The analysis of NRY data produces a tree like the one shown in figure 1. As shown the tree shows several population clusters defined by branches from a central point. The major branches are European, Middle eastern, East Asian, East European and central Asian. It is found that the Central Asian populations have high haplotype diversity. This seems to suggest that the central Asian population are the oldest in the Asian continent. This pattern of high diversity implies an early

settlement of central Asia by anatomically modern humans, perhaps 40,000-50,000 years ago followed by subsequent migrations into Europe, America, and India. It is also found that the ancestor of major European haplotype and Native American haplotype is found at polymorphic frequencies mainly in central Asia suggesting that the source of both migrations is central Asia. The distribution of a certain haplotype M17 among people of Europe and central/southern Asia suggests an ancient migration of people originating in southern Russia/Ukraine. These people are thought to have spoken an early Indo-European language. These genetic inferences are further substantiated by the following facts. (1) The inferred age of a particular haplotype M45(45,000 years) coincides with the first appearance of modern humans in southern Siberia also suggesting a migration of central Asian people to America. (2) The genetic data of migration from Russia to other parts is also substantiated by the observations that domestication of the horses starts around this period. It is also consistent with the present day distribution of languages among these people.

In the method using the SNP region of chromosome 21, the haplotype distribution also produces a tree pattern for migration. The majority of humans are parceled into 6 haplotypes Ht1-Ht6. It is found that the overall diversity is highest in Africa in whom most of the haplotypes are represented. It is also found that there are certain haplotypes not present in Africans but present in other continents suggesting that these mutations happened after the migration out of Africa. This particular range of haplotype distribution suggests that Oceania was settled after Africa(inferred from the fact that Oceanians population has the next largest haplotype distribution after the Africans). The presence of high frequence of a particular haplotype Ht6 in East Asia and not in other populations suggests that this migration was independent of other migrations. Its less frequent distribution also present in South Asia suggests the possibility of migration route as Africa-South East Asia and East Asia. The absence of a particular haplotype in European and South Asia population suggests that these regions might have been populated by a separate migration. All this points to three separate migrations out of Africa. One to Oceania, one to Asia and America and one to Europe.

The linkage disequilibrium study points out that in European populations the correlation between the SNP's last over 60 kilo bases, see figure 2. On the other hand in the Nigerian population the correlation rapidly falls of with distance(correlation of 5 kilo bases). It is suggested that LD around an allele arises because of selection or population history -a small population size, genetic drift or population mixture- and decays owing to recombination, which breaks down ancestral haplotypes. The extent of LD decrease is in proportion to the number of generations since the LD generating event. The simplest explanation for the observed long range LD is that the population under study experienced a bottleneck: a period when the population was so small that a few ancestral haplotypes gave rise to most of the haplotypes that exist today. This data is consistent with a computer simulation of the population experiencing a bottleneck which is modeled by inbreeding. This along with the fact that LD is small for African population suggests that after the European population diverged from African population they experienced a founder effect of high inbreeding.

In the Viral case, the constructed phylogenetic tree based on the most conserved region of this virus has Africa as its root indicating an African origin of the Virus. This data along with the mutation rate suggest a divergence time of at least 10,000 years for this virus and makes this a good pre historic marker of human migration. It is also found that the isolates from South east Asia is most closely related to African variety suggesting an early migration of humans from Africa to south east Asia.

Each of the above methods have their limitations and advantages, and each is able to give different information or to verify the information provided by another method. The limitations of the NRY analysis are that the population size of the Y chromosome is small and there is further confounding effects of sexual and/or natural selection. Also it is to be noted that this study wholly excludes maternal effects. The other methods also have assumptions along these lines. The Chromosome method also assumes that the effect of random mutation at this region is negligible. Note that the LD method has an advantage over other methods in that it uses simultaneous assessment of LD at multiple regions of the genome than the other methods which uses properties of single regions. This and extensions of this approach may have greater sensitivity to certain aspects of history. The method of using virus at present does not show to have any specific advantages over using human genome information. Another point of fundamental importance is that all these methods use statistical reasoning and the predictions and hypothesis from different methods have the associated statistical uncertainities. It seems that if one has to be sure of the conclusions one should varify it by more than one method and should take care to note the statistical procedure used in reaching a particular conclusion.

With the availability of human genome data it is to be expected that this kind of historical reconstruction will progress further. The LD method is a precursor to this in the sense that it seeks to unravel historical information coded as correlations between genetic regions. The progress in this field is contingent to the identification of better and better genetic regions satisfying the requirements of ideal historical markers and availability of improved statistical methods which can give more reliable and more detailed informations. Another possible application of these methods is to the determination of time period of a particular mutation event. Till now the methods used are imperfect in the sense that it makes certain assumptions about the nature and rate of mutation which is proposed to hold at all times including the past. The known chronology of pre-historic event can be used as better tests for mutation assumptions and to get information about the nature of particular mutation events.

References

[1] Lahr, M.M. (1996) *The Evolution of Human Diversity: A study of cranial variation*

[2] Cavalli-Sforza, L.L, Menozzi, P. and Piazza, A. (1994) *The History and Geography of Human Genes*

[3] Jin et al. *Proc. Natl. Acad. Sci. (USA) 96,3796 1999*

[4] Spencer et al. *Proc. Natl. Acad. Sci. (USA) 98,10244 2001*

[5] Reich et al. *Nature 411 199 2001*

[6] Pavesi A. *J. Mol. Evol. 53 104 2001*