APPENDIX 1: A BRIEF INTRODUCTION TO MOLECULAR ANTHROPOLOGY

In the following I give a very brief overview of some of the main issues involved in the use of genetic data to elucidate human prehistory. This will hopefully facilitate understanding the discussion as well as the genetic results of this study, which were published elsewhere (Pakendorf et al. 2006, 2007). For a more detailed description the reader is referred to the text book by Jobling, Hurles & Tyler-Smith (2004).

Molecular anthropology is the branch of biological anthropology that uses molecular genetic methods to study the origin, relationships, history, and migration patterns of human populations. Like other physical anthropologists, molecular anthropologists study the differences (polymorphisms) between humans. These polymorphisms exist at the molecular (protein and DNA, see below) level as much as at the physical level such as hair or skin colour, etc. Assuming an ultimately single origin of modern humans, similarities in the genetic make-up of individuals or populations may indicate a shared history and therefore relationship, whereas differences can indicate separate histories over a time frame specified by the genetic system studied.

1 The human genome

In animals and plants, the genetic information needed to build cells and run metabolic processes is stored in large molecules called deoxyribonucleic acid (DNA). DNA is present in two regions of the cell: the nucleus and the mitochondria. The DNA molecules present in the cell's nucleus are so large that at certain times during the cell cycle they can be seen under an ordinary light microscope as linear structures: the chromosomes. (That is, each chromosome is a DNA molecule.) Humans have 46 chromosomes in each cell (the exception being germ cells, i.e. ova and sperm, which contain only half the number of chromosomes), of which 44 form 22 pairs of homologous chromosomes. It used to be thought that these homologous chromosomes are identical in that they contain the same genes in the same linear order, and that they differ from each other only in the specific allele present at each locus. However, recent investigations have shown that even phenotypically normal, healthy humans differ by large-scale rearrangements of DNA such as inversions, insertions, and deletions (Iafrate et al. 2004; Tuzun et al. 2005; Conrad et al. 2006, *inter alia*). The 22 pairs of homologous chromosomes are termed autosomes, and are numbered 1 through 22. The remaining two chromosomes are the sex chromosomes, which in women form a homologous pair, whereas in men they are not homologous.

The sex chromosomes present in women are called X-chromosomes, while men have one X and one Y-chromosome. Every individual receives one set of 22 autosomes plus a sex chromosome from his or her father, and one set of 22 autosomes plus an X-chromosome from his or her mother.

In addition to the chromosomes present in the nucleus of the cell, DNA molecules are also found in the cell's energy-producing bodies, the mitochondria. This mitochondrial DNA (mtDNA) differs from the nuclear DNA in several ways. Firstly, it is much shorter (on average, nuclear chromosomes are approximately 7,800 times longer than mtDNA) and it is circular, instead of linear. Secondly, whereas the nuclear DNA is present as different pairs of homologous chromosomes, mtDNA is present in practically identical[1] copies. Furthermore, the number of molecules differs between nuclear DNA and mtDNA: there are only 23 different pairs of chromosomes (i.e. 46 molecules altogether) in each cell's nucleus, while there are hundreds of mtDNA molecules distributed within the several mitochondria in each cell. Lastly, mtDNA is passed on to the next generation only by women, i.e. it is inherited only in the maternal line (see below).

2 DNA in the germ line

It may be advisable to define some terms at this point: gene, locus and allele. A gene is a stretch of DNA that contains the information necessary to synthesize a protein (although occasionally the product encoded by a gene is not a protein, but a regulatory molecule called RNA). In general, genes consist of alternating sequences of coding DNA that can be translated into the sequence of amino acids making up the protein, and non-coding DNA, which is removed before DNA synthesis. A locus is a particular stretch of DNA and/or the product encoded by such a stretch of DNA. Since molecular anthropologists are interested in variation, the loci they study are polymorphic, i.e. at least two variants can be detected in human populations. Each of these variants is an allele at that specific locus.

In sexually reproducing species such as humans, germ cells (sperm and ova) differ from the other cells of the body in the number of chromosomes they carry. In normal cells, which carry two homologous sets of 23 chromosomes, the entire genome is present in duplicate – normal cells are therefore said to be diploid. In contrast, germ cells carry only one set of chromosomes, and are therefore said to be

[1] The mtDNA molecules within one individual may differ from each other in single nucleotides. This is termed heteroplasmy, and it arises through new mutations occurring on one molecule which gets replicated and can so spread within the cell. If an ovum contains heteroplasmic mtDNA, the arising offspring may be heteroplasmic.

haploid. One very important feature of germ cell production is recombination. Before the cell division, the two homologous chromosomes of each pair align. The two molecules physically cross each other, and sometimes during this process of crossing-over they break at the contact points. Frequently, in the fusion of such breakpoints two pieces from the maternal and paternal chromosome, respectively, are joined, resulting in an exchange of DNA stretches between two homologous chromosomes. Recombination shuffles the alleles inherited from the mother and those inherited from the father, resulting in a new mix of alleles on each chromosome. In addition, the resulting germ cell receives a random mix of chromosomes inherited from the mother and chromosomes inherited from the father. This process ensures enormous variation among an individual's offspring.

There are, however, exceptions to this biparental inheritance of DNA with recombination. First of all, mtDNA does not undergo recombination; that is, the genes carried on the mtDNA do not get shuffled (Ingman et al. 2000; Jorde and Bamshad 2000; Kumar et al. 2000; Behar et al. 2007). Furthermore, sperm carry very little mtDNA and the little that does enter the ovum at conception apparently gets eliminated within a few days (cf. references in Pakendorf & Stoneking 2005). This means that although every individual carries hundreds of mtDNA molecules in his or her cells, these are inherited solely from the mother[2].

On the other hand, the Y-chromosome in men's cells has no homologous counterpart. During cell division and germ-cell production the tips of the X and Y-chromosome align and recombine, but the major part of the Y-chromosome (termed NRPY – non-recombining portion of the Y-chromosome) does not undergo recombination (Jobling & Tyler-Smith 2003). In contrast to mtDNA which is inherited only through the mother, the Y-chromosome is passed on only from fathers to sons.

---

[2] A recent case study of a patient with exercise intolerance actually found that 90% of the mtDNA in his muscle tissue was inherited from his father, not his mother (Schwartz & Vissing 2002). However, since the other tissue samples tested (skin, blood, and hair) contained only maternal mtDNA, this seems to have been due to an accidental non-elimination of the paternal mitochondria during fertilization, which then ended up in certain cell lines of the embryo. A later study reported that the mtDNA molecules of maternal and paternal origin in this patient's muscle tissue had undergone recombination (Kraytsberg et al. 2004), demonstrating that this is possible. However, since in general the mtDNA molecules within a single individual are of uniparental origin, recombination between these will have no effects, and the non-recombining maternal inheritance of mtDNA can still be assumed for purposes of population studies.

The differences in inheritance between autosomes, sex chromosomes, and mtDNA lead to differences in the way polymorphisms on these molecules will behave in a population. In effect, when a man and a woman reproduce, there are four copies of every autosome that have a chance of being passed on to the next generation, as against three copies of the X-chromosome (two in the woman, one in the man), and only one copy each of the Y-chromosome and the mtDNA (which are passed on only from the man and the woman, respectively). This means that there are four times fewer Y-chromosomes and mtDNA molecules than autosomes participating in reproduction, and thereby in the propagation of particular polymorphisms. Furthermore, not every individual in a population participates in reproduction: some have passed their reproductive phase, some are as yet sexually immature, and not every sexually mature individual will find a mate or be able to reproduce. This means that the actual number of reproducing individuals, which is termed effective population size (N_e), is smaller than the census population size. One can therefore say that the effective population size of the Y-chromosome and the mtDNA is four times smaller than that of the autosomes, and three times smaller than that of the X-chromosome (Jobling, Hurles & Tyler-Smith 2004: 134). This has implications for population genetic analysis, since chance events and population movements have a stronger effect on smaller populations than on larger ones. Chance events will therefore effect the variation on the Y-chromosome and the mtDNA more than autosomal variation.

3 mtDNA

One of the most frequently studied stretches of DNA in molecular anthropology is the mitochondrial DNA (mtDNA). This molecule has a number of advantages for studies of population history (for a more detailed account see Pakendorf & Stoneking 2005). One of these is its high frequency in the cell, making studies possible even when only minute quantities of DNA are available. A further important advantage of mtDNA is that, although there are multiple copies of individual molecules present in every cell, these are all identical, i.e. one has to deal with only one sequence of nucleotides. (That is, mtDNA is in effect haploid as opposed to the diploid nuclear genome, where every autosomal gene is present in two copies; see above). Furthermore, as mentioned above, mtDNA does not undergo recombination, as does autosomal DNA, so that mutations remain in the context in which they arise (cf. section 5 below); this allows one to reconstruct genealogies of mutations. A further useful feature is that mtDNA is inherited solely in the maternal line, as described above. Therefore, studies of mtDNA variation illuminate

specifically the maternal history of a population, and, as recent studies have shown, this can differ from the paternal history (Passarino et al. 1998; Helgason et al. 2000a, b; Oota et al. 2001; Goodacre et al. 2005; Nasidze et al. 2006). Also, the smaller effective population size of the mtDNA as opposed to autosomal DNA (see above) means that processes affecting the genetic variation at this locus, such as population migrations or strong reductions in population size, will have much stronger effects and will therefore be more easily detectable.

There are two different approaches to study mtDNA variation. One is to establish the actual sequence of bases in a particular stretch of the molecule. Very often, molecular anthropologists focus on sequences of the so-called hypervariable region I (HVR I) of mtDNA. This is a 360 basepair (bp) long stretch of non-coding DNA that has a high rate of mutation. Sequencing this region means that every single mutation, and therefore even minute differences between individuals and/or populations, can be detected. Alternatively, molecular anthropologists study known polymorphisms dispersed over the whole mtDNA molecule. Although this approach does not detect every polymorphic site, it has the advantage of covering the entire mtDNA molecule, instead of focussing on one small stretch of DNA. With the advent of relatively cheap sequencing technology it is becoming more feasible to combine both approaches by sequencing the entire mtDNA molecule in several individuals and even whole populations (Ingman et al. 2000; Herrnstadt et al. 2002; Reidla et al. 2003; Fraumene et al. 2006; Olivieri et al. 2006), thus combining the advantages of both approaches.

4 The Y-chromosome

The counterpart of mtDNA is the Y-chromosome, which has been the focus of molecular anthropological studies in recent years (cf. Jobling & Tyler-Smith 2003 for a more detailed overview). The Y-chromosome is inherited only in the paternal line, so that it illuminates the history of a population's men. Since the Y-chromosome is present in only one copy in the cell nuclei of men (who have one X- and one Y-chromosome), it is haploid, like mtDNA. The major portion of the Y-chromosome also does not undergo recombination, and therefore mutations can be traced back in time. Studies of Y-chromosomal variation in humans focus mainly on two kinds of polymorphisms: STRs and SNPs. STRs (short tandem repeats; also termed microsatellites) are stretches of DNA consisting of repeated short sequences of DNA, for example dinucleotide repeats such as $(CA)_n$, or the tetranucleotide repeat $(CAGT)_n$. They vary in the number of repeats present in different individuals – for example, some individuals may contain 16 copies of a particular repeat

sequence, while others may have 17, 18, or more copies. STRs have a high mutation rate and can therefore be of importance for fine-grained studies of population history.

As the name suggests, SNPs (single nucleotide polymorphisms) are polymorphisms at which a single base in a certain stretch of DNA has been substituted for another. The mutation rate at these sites is so low that it is assumed that a specific substitution has occurred only once in human prehistory. This implies that all individuals (in the case of Y-chromosomal SNPs, all men) carrying a substitution at a specific site are genetically related. Therefore, SNPs are a powerful tool for establishing relationships of human populations and following paths of migration.

5 Haplogroups and haplotypes

Mutations arising on a DNA molecule in any individual in a population may spread through the population from generation to generation by reproduction. Since mtDNA and the Y-chromosome do not undergo recombination, as do autosomes, a mutation on these types of DNA will remain in its original context within the molecule. (In autosomes, such a mutation might well be passed to a different homologous chromosome, with a different DNA context, through recombination.) With time, further mutations may arise on a mtDNA molecule or Y-chromosome carrying the original mutation, leading to several mutations being associated with each other. This is called a haplotype. A particular haplotype is defined by the researchers conducting the study in order to describe their results. Sometimes, a haplotype is found in only one individual, whereas other haplotypes may be found in several individuals.

A haplogroup, on the other hand, is a group of related haplotypes that is defined by researchers on account of a recurrent association of specific mutations (‘shared innovations’). A haplogroup is an indicator of genetic relationship, since it is assumed that the mutations defining it arose in a linear manner at different points in time on certain DNA molecules, and then spread through the population by reproduction of these molecules. Languages and language families are possible linguistic analogies to haplotypes and haplogroups, respectively: a language is defined by having a certain set of features, and a language family is a group of genealogically related languages that share certain, but not all, of the features of the individual languages.

The nomenclature of haplogroups in mtDNA has developed over time, with related groups of haplotypes being classed together as they were found in studies of

populations from around the world. Most mtDNA haplogroups are defined by SNPs in the coding region of mtDNA, but some can also be distinguished on the basis of certain HVR1 sequence types (haplotypes). Since the definition of these haplogroups proceeded in an *ad hoc* manner, the nomenclature lacks any coherent logic. Haplogroups A through G are found in Asia; of these, A to D also occur in the New World, where they make up the vast majority of mtDNA types found. Haplogroups H, I, J and K are found in Europe, with H and I present in western Asia as well. All the mtDNA variation found in Africa has been classified as (super)haplogroup L, with numbered subscripts differentiating between subgroups (L_0–L_6). Haplogroups T, V, W, and X are European haplogroups; interestingly, haplogroup X is also found in the indigenous populations of the New World. Recently, this haplogroup was also found in individuals from the Altay mountains in south Siberia, providing the first link between haplogroup X in Europe and in the Americas (Derenko et al. 2001). Finally, haplogroups Y and Z are found in restricted populations of northeastern Siberia and the Far East.

Researchers studying Y-chromosomal variation also distinguish haplogroups based on SNPs. Since the mutation rate of these SNPs is so low that they have arisen only once in human history, they are ideal markers for defining groups of related individuals: every man carrying a particular SNP is genetically related to every other man carrying this SNP. Of course, some SNPs are very widespread and therefore not of much use for studies of specific populations, whereas others are limited to certain geographical areas and can therefore provide information on the relationship of populations in these areas. One of these localized SNPs is the so-called “Tat” marker mentioned in sections 1.1.1.2 and 1.1.3.2, which is very widespread in populations of northern Eurasia and practically absent or very marginal elsewhere (Zerjal et al. 1997; Karafet et al. 1999; Lahermo et al. 1999; Rootsi et al. 2007). This may have arisen in northern China and spread throughout Siberia to Europe (Rootsi et al. 2007). After a period of nomenclatural confusion the Y-chromosomal haplogroup nomenclature has been unified (YCC 2002). Haplogroups of importance in Siberia are C, N and Q, while haplogroup O is widespread in Southeast and East Asia. Haplogroup J is found in Europeans, while haplogroup R is widespread in Europeans and Central Asians (Jobling & Tyler-Smith 2003).

The genetic distance between individuals can be measured directly by counting the number of mutations that have accumulated since they shared a common ancestor. However, if new mutations arising on diverging lineages of human populations were the only means of analyzing the prehistory of these populations, the time depth that could be investigated by molecular genetic methods would be far too early to be of interest to biological anthropologists. Fossil evidence points to an origin of anatomically modern humans in Africa between 200,000 to 100,000 years before present. In the fast-evolving HVR I of mtDNA a new mutation arises on average every 15,000 to 20,000 years. Therefore, only five to ten new mutations will have arisen on each lineage in the HVR I since the origin of modern humans – which implies that current-day populations would be too similar to permit any meaningful study of their past history. Questions of recent population history, such as the migration of the Sakha from South Siberia to the middle reaches of the Lena, could not be answered at all. The reason that molecular anthropologists continue to make claims about such recent events is that new mutations are not the only source of inferences about the past. Other random factors can shape the patterns of genetic variation much more strongly.

One important factor is the fate of the preexisting genetic variation in the ancestral population. If a population starts to split, each daughter population will receive some of the variation already present. Depending on the geographic and cultural circumstances involved, the daughter populations may each receive a slightly different subset of the initial variation. Furthermore, some of the mutations present initially in the daughter populations may disappear through random factors – individuals carrying these polymorphisms might not find a mate and therefore will not reproduce, resulting in their genetic lineage dying with them, or men might have only daughters, or women might have only sons, resulting in the termination of their Y-chromosomal and mtDNA lineages, respectively. These random processes affecting the genetic variation in a population are termed genetic drift. Genetic drift has a stronger effect in a small population than in a large one, so that demographic factors such as population size play a role in shaping genetic variation, too. For example, an allele with a frequency of 1% would be present in 1,000 individuals in a population of 100,000, whereas in a population of 1,000 this allele would be present in only ten individuals. It is more probable that ten people die young, or do not find a partner, or have children who die young, than this happening in 1,000 individuals (cf. Jobling, Hurles & Tyler-Smith 2004: 131ff).

An extreme case of genetic drift is the so-called founder event. In this event, a small group of individuals leaves the ancestral population and migrates to a

different location, for example an island. This founding group will carry only a small subset of the variation present in the entire ancestral population, and this subset will thus be present at high frequency in the resulting new population. A very good example of a founder event is the colonization of the New World, where four of the seven mtDNA haplogroups defined in Asia are present in very high frequency in the indigenous populations, whereas the other three are completely absent (Torroni et al. 1993, Stone & Stoneking 1998). Similarly, should a large number of the ancestral population rapidly die off, the survivors will also retain only a subset of the initial variation. This event is termed a bottleneck, and it, too, can lead to differences between populations that share a common ancestor. Identifying whether the prehistoric event was a bottleneck or a founder event is not easy, since the result seen in the current-day population is the same: reduced genetic variation that represents a subset of the variation present in the presumed ancestral population (Jobling, Hurles & Tyler-Smith 2004: 132f). To distinguish between a founder event and a bottleneck, data from other fields, especially archaeology, are needed.

7 Methods of inferring population relationships: Fst, MDS, and AMOVA

Population relationships can be inferred from molecular genetic data by estimating genetic distances between individuals or populations. One measure that is frequently used as an estimate of genetic distance is Fst (Jobling, Hurles & Tyler-Smith 2004: 168). Fst is the proportion of the total genetic variation[3] that is due to differences between populations. Fst is estimated by subtracting the genetic variation that exists within individual populations from the genetic variation found in the combined populations. If two populations do not differ genetically, the amount of variation present within each of them will make up all or most of the genetic variation present within the total sample, and therefore the Fst value will be close to zero (cf. the non-significant Fst value of 0.02 between Central Ėvens and Western Ėvens in Table 3 of Pakendorf et al. 2007). If, however, populations are genetically very different, the Fst value will be significantly larger than zero (even though most of the variation within the total sample will still be due to the variation existing within them; cf. the significant Fst value of 0.21 between the Iengra Evenks and the Stony Tunguska Evenks (STE) in Table 3 of Pakendorf et al. 2007). The significance of Fst values can be estimated by permutation, which is of great

[3] Technically, Fst is the proportion of the variance that is due to differences between populations. However, for the purposes of this introduction, genetic variance and genetic variation are equivalent.

advantage in data interpretation, since it is allows one to distinguish small values that are only due to chance from small values that are truly indicative of shared history. In order to test the significance of Fst values by permutation, all the haplotypes of the total sample are pooled, and then randomly allocated to "populations" equivalent in size to the original populations in the comparison. The Fst value is calculated for this random sample of "populations", and then the procedure is repeated; in general, to assess the significance of Fst values one performs 10,000 permutations. At the end of the procedure, the real, observed Fst values can be compared with the randomly generated ones to assess whether they, too, are due to chance or whether they are significantly larger than these random values. The measure of Fst was first developed by population geneticists for classical allele frequency data, but it has since been expanded to take account of the new data being generated nowadays. Measures related to Fst are Rst, which is adapted specifically to the singular mode of mutation of STRs, and Φst, which is specifically adapted to molecular data – however, this latter is often called Fst in the literature, such as in this thesis and in the two articles in which the genetic results were published (Pakendorf et al. 2006, 2007).

Fst values are usually presented in tables (e.g. Table 3 in Pakendorf et al. 2007) and are often depicted either as trees or with the help of multidimensional scaling analysis (MDS). Since the Fst values estimate the genetic distances between populations, such trees and MDS plots are a way of depicting the relationships between the populations (cf. Figure 4 in Pakendorf et al. 2006, and Figure 2 in Pakendorf et al. 2007). MDS is a multivariate method that attempts to arrange the objects of study (e.g. population samples of mtDNA sequences) in geographical space in such a way that the distances between the objects are reproduced as accurately as possible (STATISTICA for Windows 1995: 3237). The advantages of MDS plots over trees is that trees will always join all the populations in a sample, whether or not this reflects the true genetic relationships. This can falsely create the impression of a genetic relationship where none exists. MDS plots are not constrained in this manner, and isolated populations will remain at a distance from all the other populations in the sample without being arbitrarily linked to any of them. Furthermore, trees are in effect unidimensional, while MDS analyses utilize more of the information present in the data by extracting several dimensions (although it is impossible to visualize more than three dimensions at once).

An alternative method to calculating genetic distances or Fst values and visualizing them in trees or MDS plots is to analyze the relationship of the actual sequences or STR haplotypes by constructing networks among them (cf. Figure 2a and 3a in Pakendorf et al. 2006). Since in a network every sequence or haplotype is

linked to those sequences or haplotypes that are most closely related to it, all possible trees are shown (Jobling, Hurles & Tyler-Smith 2004: 175ff).

A method that is conceptually based on Fst analysis is AMOVA (Analysis of Molecular Variance), which allows one to examine the effect of external factors (such as geography or language) on the genetic differences between populations. AMOVA estimates the proportion of variation at different levels of population groupings. One level examined is the genetic variation between populations within a group of populations defined by the researcher, based for example on linguistic affiliation or geographic source of the populations concerned. The second level is the genetic variation that exists between two or more such defined groups, while the third level is the variation present within the individual populations of the sample. If the groups defined by the researcher fit well with the underlying genetic variation, the genetic variation among the populations **within** the groups will be small, while the genetic variation **between** the different groups will be large. For example, one could group a sample of populations according to their linguistic affiliation; if linguistic relationship reflected genetic relationship, the populations grouped together under each language family label should be genetically homogenous. In such a case, the proportion of variance among the populations within each group (language family) would be small, while the proportion of variance between the individual groups (language families) would be large. If, on the other hand, linguistic relationship should NOT reflect genetic relationship, for instance because of prehistoric language shift, then the proportion of variance within the language family groups would be large, while the proportion of variance between the groups would be relatively small (cf. Table 5 in Pakendorf et al. 2006, where the proportion of variance within the linguistically defined groups is 4.06%, as opposed to only 1.94% of the variance being due to between-group variation).