Molecular Markers in Anthropological Genetic Studies
Rohina C. Rubicz, Phillip E. Melton and Michael H. Crawford

Laboratory of Biological Anthropology, University of Kansas, Lawrence, KS

Introduction

In 1973, a chapter entitled 'The use of genetic markers of the blood in the study of the evolution of human populations', was published in the first volume that attempted to synthesize the field of anthropological genetics (Crawford, 1973). This chapter defined genetic markers as 'discrete segregating, genetic traits which can be used to characterize populations by virtue of their presence, absence, or high frequency in some populations and low frequency in others' (Crawford, 1973: 38). This definition similarly applies to molecular markers, which are segregating regions of DNA, present in some populations but absent or infrequent in others. The 1973 chapter summarized the available genetic markers of the blood that could be used for the measurement of evolutionary processes and the characterization of human population structure. The list of available polymorphic loci included 16 blood groups, 11 red blood cell proteins, 10 serum proteins and 3 white cell and platelet systems. These 'riches' of available variation of the blood followed 70 years of research on the blood group systems (since Karl Landsteiner’s original work in 1900), and Oliver Smithies (1955) development of zone electrophoresis for the separation of specific proteins from mixtures such as the serum of the blood (Landsteiner and Levine, 1927). At the time the first volume in anthropological genetics was compiled, the physiological functions of blood groups were unknown, other than their involvement in blood transfusion and some suspect statistical associations with disease. The role of Duffy blood group as a resistance factor to Plasmodium vivax in Africans was not known

This research was sponsored by NSF grants OPP-990590 and OPP-0327676.
Molecular markers in anthropological genetic studies

until 1976 (Miller et al., 1976). In 1993, Horuk et al. demonstrated that Duffy blood group antigen was a chemokine receptor that the malarial parasite utilized for its entry into the erythrocyte. The function of the Rhesus blood group system and the homology between the mep gene in yeast and the ammonia transport system across cell membranes was unknown until the molecular genetic research by Marini and Urrestarazu in 1997. In the 1970s, from an evolutionary perspective, we knew little about genetic markers and their functions. Our genome was thought to consist of 100,000 genes with 30,000 estimated to be polymorphic with fewer than 100 loci had been adequately documented.

Methodological developments

The past three decades have seen a number of major technological developments that graduate students in the 1960s could have only dreamed about. Variation in primary gene products (proteins) or secondary gene products (such as blood groups) were used as proxies for the DNA that synthesized them. The technological changes that eventually resulted in the direct manipulation of DNA included:

(1) Rapid methods of DNA extraction. The first isolation of DNA occurred in the late 1860s by Friedrich Miescher, who was attempting to investigate the characteristics of proteins in pus cells. He observed that alkaline cellular extracts, when neutralized, yielded a precipitate that he termed 'nuclein'. Classical techniques for extracting DNA required from 3 to 24 hours to complete and the use of expensive equipment and caustic and toxic chemicals. Early extraction techniques also required large quantities of cellular materials, with placenta and blood being the preferred tissues. Now, modern protocols bundled into extraction kits can purify DNA in as little as 30 minutes and do not require toxic chemicals. Most of these methods are based on cell lysis, protein degradation with the destruction of the nucleases, nucleic acid precipitation and fractionation.

(2) The discovery that restriction enzymes produced by bacteria cleaved DNA at or after specific sequences and could be applied to solving problems in molecular genetics won the 1978 Nobel Prize in Physiology or Medicine for Hamilton O. Smith and Dan Nathan. This discovery enabled early molecular specialists to identify mutations and variants in specific regions of the genome. This was a less labour-intensive method than sequencing for the characterization of restriction fragment length polymorphisms (RFLPs) in mitochondrial and nuclear DNA.

(3) DNA hybridization techniques (annealing two homologous DNA strands) that allowed comparisons of DNA between different species. This method compares single strands of
two different DNA molecules that are re-associated and their thermal stability is ascertained. The original experiments by Sibley and Ahlquist (1984) demonstrated that chimpanzees and humans were more similar than either species was to gorillas and placed the chimpanzees as our closest phylogenetic relative.

(4) Polymerase Chain Reaction (PCR) revolutionized molecular genetics by making copies of DNA sequences through flanking a particular region with primers, denaturing the DNA, and annealing to the target sequence. The method of alternate heating and cooling allows the synthesis of specific DNA regions in geometric progression. Development of the PCR method and the thermocycler in the late 1980s won for Kary Mullis the 1993 Nobel Prize in Chemistry. This methodological breakthrough has had profound effects on anthropological genetics and forensic sciences. Researchers in these fields deal with minute quantities of DNA that must be amplified for either genotypic characterization or sequencing of regions of the genome.

(5) Automated DNA sequencing, followed by the development of high throughput sequencing in the late 1990s resulted in the rapid characterization of the human genome. The earliest methods of sequencing were extremely time intensive and were based on either chemical identification of the purines and pyrimidines or the use of dideoxyriboside chain terminators (Maxam and Gilbert, 1977; Sanger et al., 1977). Both methods required the identification of specific DNA fragments on polyacrylamide gels followed by radio-labelling (see Devor, Chapter 10 of this volume for a discussion of the development of sequencing methodologies). The high throughput methods using fluorescence dye-terminators and capillary electrophoresis sped up the sequencing exponentially and made possible the completion of a draft of the human genome sequences in 2001.

The remainder of this chapter provides: (1) a description of the available DNA markers that can be utilized in anthropological genetic investigations; (2) a discussion of the most common and informative analytical tools that can be applied to molecular marker distributions in populations to answer evolutionary and historical questions; (3) the applications of specific DNA markers and analytical techniques to the study of evolutionary processes.

Molecular markers

In humans, DNA is packaged into 22 pairs of autosomes and a pair of sex chromosomes (XX or XY), but is also present outside the nucleus in the form of mitochondrial DNA (mtDNA). Based on function,
nuclear DNA can be subdivided into coding and non-coding regions. Coding DNA, i.e. genes or exons, are defined as sequences that carry instructions for synthesizing proteins. Non-coding regions of the DNA make up the largest portion of our genome with an estimated 98.5% of the total nuclear DNA. They are not under the functional constraints of genes, therefore allowing them to exhibit greater variation, and are generally considered to be selectively neutral. Previously described as ‘junk’ DNA, it is now evident that some non-coding regions may function in providing structural support for DNA molecules and may be involved in gene regulation.

Several types of polymorphisms are present in the human genome. The most basic of these is the substitution of one base for another, called a single nucleotide polymorphism (SNP). Insertions and deletions, ‘indels’, of single bases are sometimes included in the first category, although the mechanism by which they arise differs. Larger indels also occur throughout the genome. Another category, that of repetitive DNA, comprises approximately 45% of the genome (Lander et al., 2001) and includes sequences repeated in tandem arrays, such as micro- and minisatellites, and retroelements (DNA sequences that have been inserted into the genome through reverse transcriptase).

**Autosomal markers**

Anthropological genetics studies have utilized molecular markers characterizing polymorphisms occurring at various locations throughout the genome to investigate questions concerning the history of a population, population structure, and events such as migration and gene flow, including the time frame in which they occurred. Studies using autosomal markers have an advantage over those using Y chromosome and mtDNA markers in that they sample a larger portion of the gene pool and may be more representative of a population as a whole. Autosomal polymorphisms are biparentally transmitted, therefore they are not limited to either the paternal (in the case of the Y chromosome) or the maternal (in the case of mtDNA) side, but rather can provide information about both sexes. Alleles at different loci may undergo recombination and assort independently of one another as they are passed down each generation. This reshuffling generates novel combinations of genetic material, although depending on the particular research question, it is not always a desirable characteristic. The effective population size or $N_e$, an estimate of the breeding size of a population, is largest for the autosomal loci since they are present in two copies for both males and females. In contrast, there are three-quarters the number of X chromosomal loci (2 X’s for each female, and 1 X for each male), quarter the number of Y chromosomal loci in the same population assuming equal numbers of males and females, and quarter the
number of mtDNA loci. Markers present on the X chromosome and pseudo-autosomal regions of the Y chromosome (small areas at the tips that recombine with the X chromosome) are otherwise similar to autosomal markers in their analysis and potential for resolving anthropological questions such as population history and relationships between populations.

Of the estimated 32,000 genes present in the human genome, the majority are located on the autosomes, while the X chromosome has approximately 1,500 genes, and the Y contains approximately 78 (Venter et al., 2001; Skaletsky et al., 2003). These coding regions are under functional constraint, since mutations that disrupt protein synthesis (such as indels that shift the DNA reading frame) can result in genetic diseases. Sometimes these mutations are associated with lowered fitness or even death of the individual. In these cases the variant is likely to be quickly eliminated from the population. Some mutations may have no effect on the phenotype, for example a base substitution in the third position of a codon often codes for the same amino acid. Still other mutations may produce variants of a gene that are advantageous to a human population living in a particular environment.

Collagen genes (COLIA2)

One example of an autosomal coding marker found to be informative for human phylogenetic research is the human α2 (1) collagen gene (COLIA2) used by Mitchell et al. (1999). COLIA2 is one of two genes that code for peptides in type 1 collagen (a component of skin, bone, blood vessels, ligaments, and dentin, among other tissues). In Mitchell et al.'s (1999) study, the worldwide distribution of a 38 base pair deletion in this gene among human populations (which appears to be neutral) was characterized in order to test the recent out-of-Africa model. This model, supported by previous molecular studies (Vigilant et al., 1991; Stoneking and Soodyall, 1996; Hammer et al., 1997, 1998), states that modern Homo sapiens originated in sub-Saharan Africa and rapidly spread throughout the rest of the world, replacing earlier hominid forms. The collagen deletion was determined to be present in high frequencies in non-African populations, but was completely absent in sub-Saharan groups. These results indicate the marker likely arose just before or shortly after modern humans left Africa, but before their spread throughout the rest of the world. A similar distribution as found in the collagen deletion was noted for the immunoglobulin GM locus. The GM*A,X,G haplotype is absent in sub-Saharan Africa but occurs worldwide in all other human populations. Thus, the population out-of-Africa experienced at least two distinct mutations before it spread throughout the world. Mitchell et al.'s (1999) study lends further support to the recent out-of-Africa model, and demonstrates that COLIA2 is a useful marker.
in reconstructing phylogenetic relationships between human populations.

**Single nucleotide polymorphisms (SNPs)**

A variety of markers located in non-coding autosomal regions are useful for anthropological genetics studies. Of these, single nucleotide polymorphisms (SNPs) tend to evolve at a slow rate of approximately $2.3 \times 10^{-8}$ (Nachman and Crowell, 2000). These simple base substitutions may result from the action of mutagenic agents, such as radiation, or be due to nucleotide misincorporation during DNA replication. Transitions, where a purine is substituted for another purine (i.e. $A \rightarrow G$) or pyrimidine for another pyrimidine (i.e. $T \rightarrow C$), occur at a higher rate than transversions, where there is a change in nucleotide class (i.e. $A \rightarrow T$). SNPs (including deletions or insertions of single bases) can be identified through RFLP analysis, whereby the target DNA is amplified by PCR, digested by restriction enzymes, electrophoresed, and scored. A single nucleotide change in the DNA sequence recognized by the enzyme will prevent its cleavage. While this method is widely used for the detection of SNPs, direct sequencing of the nucleotides is also common. The low mutation rate of these markers makes it unlikely that they have reoccurred during the evolution of modern humans, and therefore individuals sharing the same marker can be assumed to share common ancestry, or be identical by descent. The ancestral state for a particular locus can be inferred by comparing the sequence to that of our closest relatives, the great apes. Humans and chimpanzees are estimated to have diverged around 5 million years ago. SNPs, also described as binary markers or unique mutational events (UMEs), may be of use for investigating phylogenetic questions of great time depth.

**Tandem repeats**

**Short tandem repeats (STRs)**

Tandemly repeated DNA sequences evolve at a faster rate than SNPs and are common throughout the human genome. Small repeat units called microsatellites or short tandem repeats (STRs) range from 1–6 base pairs (e.g. CA-, CAT-, CCG-, CAG-) with the total repeat unit usually less than 350 base pairs (Guarino et al., 1999). These occur approximately every 6 to 10 kilobases. They are believed to typically mutate according to the gain or loss of single repeat units at a time, although they may undergo larger ‘jumps’ whereby several repeats are simultaneously inserted or deleted. The mutation rate varies among STR loci, partly due to size and composition of the repetitive unit, and the number of repeats (alleles with larger numbers of repeats tend to mutate more rapidly than those with fewer repeats).
Table 6.1. Average mutation rate for DNA markers (after Jobling et al., 2004).

<table>
<thead>
<tr>
<th>DNA Marker</th>
<th>Mutation rate per locus per generation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Some expanded polymorphic microsatellites</td>
<td>&lt;10^-6</td>
</tr>
<tr>
<td>Minisatellites</td>
<td>10^-2 to 10^-1</td>
</tr>
<tr>
<td>Microsatellites</td>
<td>10^-4 to 10^-3</td>
</tr>
<tr>
<td>Some structural polymorphisms</td>
<td>10^-5 to 10^-4</td>
</tr>
<tr>
<td>Base substitutions (SNPs)</td>
<td>10^-8 to 10^-7</td>
</tr>
<tr>
<td>Retroelement insertions</td>
<td>10^-1 to 10^-10</td>
</tr>
</tbody>
</table>

STR mutation rate is estimated between 10^-4 and 10^-3 mutations per locus per generation (see Table 6.1), making them ideal for investigations into recent human history, such as the recent separation of human groups. While a number of studies have indicated that microsatellites tend to be selectively neutral, there is evidence implicating several trinucleotide repeats in human diseases. For example, 'CCG' repeat in fragile-X syndrome, 'CAG' in Huntington's Disease, and 'CTG' in myotonic dystrophy (Fu et al., 1991; The Huntington's Disease Collaborative Research Group, 1993; Brook et al., 1992). The alleles become increasingly unstable with larger numbers of repeats, until they reach a threshold at which the disease manifests.

Variable number of tandem repeat (VNTRs)

Another class of repetitive DNA sequences consists of minisatellites or variable number of tandem repeat (VNTR) polymorphisms. Some investigators use an alternative nomenclature system in which VNTRs refer to all tandem repeats; microsatellites, minisatellites, and satellites. According to the more specific definition used here, VNTRs consist of core repeat units of approximately 10 to 100 base pairs, strung together up to 1,000 base pairs in length. They are usually GC-rich, and exhibit sequence variation in addition to variability in the number of repeats. VNTR mutation rate is higher than that of STRs, at approximately 10^-2 to 10^-1 per locus per generation (see Table 6.1), and their mutation process is more complex, with a bias toward repeat gains over losses. VNTRs exhibit a great amount of variation between individuals which makes them particularly useful for forensic studies. Their application to individual identification, using a method called 'DNA fingerprinting', such as for paternity testing, is described in greater detail by Schanfield, Chapter 9 of this volume.

VNTRs have been successfully used in anthropological genetic research to explore the Siberian origins of Native American populations. In studies by the University of Kansas research group (1989–1995), five VNTR loci (D7S104, D11S129, D18S17, D20S15,  


and D21S112) were used to characterize the population structures and genetic affinities of Siberian groups (McComb et al., 1995, 1996). Siberia is of evolutionary importance to the prehistory of human populations as a crossroads between Europe, Asia, and the Americas. In the first season of fieldwork, a comparison was made between two Evenki villages, Surinda and Poligus, and the Ket village of Sulamai (McComb et al., 1995). The Evenki are Tungusic-speaking reindeer herders who are widely dispersed throughout the taiga of central Siberia and are relatively isolated. The Kets subsist on hunting and fishing, speak a language unrelated to the three major linguistic phyla of Siberia (Altaic, Uralic, and Paleoasiatic), and are of unknown origin. VNTR frequencies were significantly different between Sulamai and Surinda at the D11S129 locus. Sulamai was shown to have experienced more gene flow from Russians, a finding in agreement with the available ethnographic information, and Poligus and Surinda were relatively isolated and experienced little gene flow from outside groups. Overall, the VNTR data demonstrated the distinctiveness of the Kets and Evenki. In a second year of field investigations, a Kizhi population from Gorno-Altai was compared to the Kets and Evenki (McComb et al., 1996). The Kizhi of the Gorno-Altai are a group of Turkic-speaking pastoralists, previously shown to be genetically more diverse than other Siberians and they apparently exhibited the so-called Asian-specific 9 bp deletion in mtDNA. The presence of the 9-bp deletion, which defined the B mtDNA haplogroup, made the Gorno-Altai unique in Siberia with the presence of A, B, C, and D haplogroups (Shields et al., 1993). More recently the Altai became of great interest to anthropological geneticists because of the reported presence of haplogroup X, shared with Native American populations (Derenko et al., 2003). The VNTR data confirmed that the Altai were indeed genetically distinct from the other Siberian groups, although not to as great an extent as previously indicated. In both studies, the Siberians were shown to cluster together when compared to outside populations, indicating their relative genetic homogeneity, and they were closer to Native Americans than to other American populations (African Americans and European Americans) reflecting their recent common evolutionary history. Thus, VNTR markers (DNA fingerprints) can be useful for discriminating between populations, measuring population affinities, and examining recent events such as human migrations.

Telomeric arrays and satellites
Telomeric arrays and satellites are repeat polymorphisms that have not been widely used in anthropological genetics. Telomeric arrays are DNA-protein structures with tandem repeats, which are located at the ends of the chromosomes, and are sometimes classified as minisatellites. They play a functional role in preventing chromosomes from fusing together and preventing chromosomal degradation. Satellites are very large tandem repeats of hundreds of
Y CHROMOSOME DNA MARKERS

thousands to millions of bases long. Some have been shown to provide structural support, such as the alpha satellite which forms part of the centromeres. Satellites are difficult to work with because of their tremendous size.

Retroelements
Retroelements are DNA elements that transpose into genomic locations after their transcription to RNA from an active genomic copy, and then reverse transcription into DNA. One example is the Alu family, consisting of short interspersed repetitive units (SINEs) which are only found among primates. An Alu consists of two approximately 150 base pair units comprising a highly repetitive sequence of around 300 nucleotides, which may be randomly inserted hundreds to thousands of times. In humans there are an estimated 500,000 copies per haploid genome (Deininger and Batzer, 1993; Novick et al., 1996). Of these, a few thousand are thought to be polymorphic in humans, and therefore useful in anthropological genetics studies. Insertion events of these elements are rare, and once inserted, they are extremely stable. The ancestral state of any Alu repeat is its absence, while its presence is considered to be the derived state. There is no known mechanism for the complete removal of an Alu from its insertion site. Other retroelements include LINEs (long interspersed nuclear elements), such as the L1 element, and HERVs (human endogenous retroviruses). These retroelements also show promise for understanding the etiology of certain diseases, as well as for phylogenetic studies of human populations.

Randomly amplified polymorphic DNA (RAPD)
Another type of marker, randomly amplified polymorphic DNA (RAPD), will only be mentioned briefly here because it has had limited use in studies of human diversity. RAPDs are generated by using short primers (8–12 bases) to randomly amplify segments of DNA. Variation is seen in the presence or absence of bands, or in different lengths of bands. Although early research on humans using RAPDs was promising in that phylogenetic reconstruction of populations of known histories gave accurate results (McComb, 1999; Melvin, 2001), this method has fallen out of favour for human studies because of criticism that the majority of fragments are non-specific and cannot be reproduced.

Y chromosome DNA markers

The Y chromosome (see Figure 6.1), composed of large amounts of chromatin and few genes (related to male sex determination), is passed exclusively from father to son. Only a small portion of the Y, the pseudoautosomal region located at the tips of its chromosomal arms, recombines with the other sex chromosome, the X.
Approximately 95% of the Y chromosome does not recombine, and it is this portion, referred to as the MSY (male-specific region of the Y), that is targeted for population studies. The MSY is said to be haploid, present in only one copy, because it does not share a corresponding region on the X chromosome. Being uniparentally inherited, MSY markers have at most a quarter of the effective population size ($N_e$) of autosomal markers. This may be even further reduced because the Y chromosome is subject to genetic drift since not all males will contribute equally to the gene pool of future generations, and some Y chromosomes may be lost through stochastic processes.

Y chromosome DNA polymorphisms are useful for the phylogenetic reconstruction of populations because in the absence of recombination, any mutations that occur will be passed on to future generations. In this way, long-lasting patrilineages can be identified and traced back to a common male ancestor. At a local level, higher female than male migration rates due to cultural practices (including patrilocality, the practice of women moving to
their husbands' place of residence after marriage) may lower Y diversity in comparison to that of mtDNA (Salem et al., 1996; Oota et al., 2001). Although some studies indicated this pattern extended to a global scale (Seielstad et al., 1998; Romualdi et al., 2002) this no longer appears to be the case, as it is likely that longer-distance migrations were more often made by males (Wilder et al., 2004).

The most widely used MSY polymorphisms for phylogenetic studies are binary markers and STRs. The binary markers, consisting mainly of SNPs and indels, evolve slowly and are considered to be unique mutational events (UMEs). They are used to define the major Y chromosome lineages, or haplogroups. Recently, the haplogroup nomenclature was revised in order to construct a single system that reflects the phylogenetic relationship between haplogroups and has the flexibility for the incorporation of new mutations (Y chromosome consortium 2002). Previous to this, seven different systems of nomenclature were in use. Y chromosome STRs are highly polymorphic and are used to characterize variation within the haplogroups. A Y specific minisatellite (MSY1) is even more variable and has been similarly used by researchers (Jobling et al., 1998; Brion et al., 2003).

The characterization of new Y chromosomal markers in populations worldwide has contributed to their increasing usefulness for anthropological genetic studies, which until recently lagged behind mtDNA studies (see the chapters by Arredi et al., Tishkoff and Gonder, and Salzano in their application of Y chromosome markers to the peopling of Africa, Europe and the Americas).

Mitochondrial DNA markers

The other non-recombining portion of the human genome is the mitochondrial DNA (mtDNA). MtDNA is a double-stranded, circular molecule believed to be of bacterial origin (Margulis 1981), and located outside the nucleus in the energy-producing mitochondria of the cell (see Figure 6.2). MtDNA is maternally inherited, meaning that it is passed from a mother to all of her children, but only her daughters will pass it on to subsequent generations. Similar to the Y chromosome, the mtDNA has an effective population size equal to a quarter that of the autosomes, and it is also subject to genetic drift. MtDNA is present in multiple copies per cell, hundreds to thousands, depending on the tissue. This characteristic is desirable for ancient DNA studies, where the material recovered from deceased individuals is often degraded (see Chapter 8 by O'Rourke on ancient DNA). There are approximately 16,569 base pairs in the mtDNA molecule, which consists of a coding region with 37 genes and two non-coding hypervariable regions (HVS-1 and HVS-2). The molecule mutates approximately ten times faster than nuclear DNA because it lacks the nuclear repair mechanisms, and the hypervariable region or D-loop
has an even faster evolutionary rate, making it useful for studies of human population history dating back 100,000 years or less.

mtDNA markers used in population studies mainly consist of RFLPs (restriction fragment length polymorphisms) and sequencing of HVS-1 and HVS-2. RFLPs (and increasingly sequencing) are used to characterize SNPs and a 9-base pair deletion located in the coding region of the molecule. These polymorphisms define the major mtDNA haplogroups, which are considered to be relatively stable. HVS-1 and HVS-2 sequences have been used to characterize diversity within the haplogroups. Because recurrent mutations occur in the hypervariable region, shared polymorphisms in this segment cannot be assumed to be identical by descent. Some researchers have even sequenced the entire mtDNA genome, providing a detailed picture of individual maternal lineages.

The Aleuts provide an informative example of the application of mtDNA markers to the study of human populations (Rubicz, 2001; Rubicz et al., 2003). The Aleuts are an indigenous Alaskan population located near the entry point to the Americas (see Figure 6.3), who represent one of the final migrations of humans into previously unoccupied territory. Like other Native Americans, they are proposed to have originated in Asia, and crossed over to Alaska by route of the Bering Land Bridge, which was exposed during the last Ice Age when sea levels were lower. Aleut mtDNA RFLP and HVS-1 sequencing variation was characterized in order to investigate their origins and role in peopling of the New World. Of the four haplogroups common among A and B and C and D native groups (see Table 6.2), the Aleuts were found to be closest to the N1 group, which is thought to be the most ancient and significant in North America.
Table 6.2 | MtDNA frequencies among indigenous Siberian and American populations (Rubicz, 2001).

<table>
<thead>
<tr>
<th>Population</th>
<th>n</th>
<th>MtDNA*A</th>
<th>mtDNA*B</th>
<th>mtDNA*C</th>
<th>mtDNA*D</th>
<th>mtDNA*OT</th>
<th>Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aleut</td>
<td>179</td>
<td>28.5%</td>
<td>0%</td>
<td>0%</td>
<td>71.5%</td>
<td>0%</td>
<td>1</td>
</tr>
<tr>
<td>Altai</td>
<td>28</td>
<td>3.6%</td>
<td>3.6%</td>
<td>35.7%</td>
<td>14.3%</td>
<td>42.8%</td>
<td>5</td>
</tr>
<tr>
<td>Asian Eskimo</td>
<td>50</td>
<td>80.0%</td>
<td>0%</td>
<td>0%</td>
<td>20.0%</td>
<td>0%</td>
<td>7</td>
</tr>
<tr>
<td>Coastal Chukchi</td>
<td>46</td>
<td>23.9%</td>
<td>0%</td>
<td>21.7%</td>
<td>8.7%</td>
<td>45.7%</td>
<td>5</td>
</tr>
<tr>
<td>Dogrib</td>
<td>154</td>
<td>90.9%</td>
<td>0%</td>
<td>2.0%</td>
<td>0%</td>
<td>7.1%</td>
<td>3</td>
</tr>
<tr>
<td>Savoonga Eskimo</td>
<td>49</td>
<td>93.9%</td>
<td>0%</td>
<td>0%</td>
<td>2.0%</td>
<td>41%</td>
<td>3</td>
</tr>
<tr>
<td>Evenki</td>
<td>51</td>
<td>3.9%</td>
<td>0%</td>
<td>84.3%</td>
<td>98%</td>
<td>2.0%</td>
<td>7</td>
</tr>
<tr>
<td>Even</td>
<td>43</td>
<td>0%</td>
<td>0%</td>
<td>58.1%</td>
<td>70%</td>
<td>34.9%</td>
<td>7</td>
</tr>
<tr>
<td>Haida</td>
<td>25</td>
<td>96.6%</td>
<td>0%</td>
<td>0%</td>
<td>3.4%</td>
<td>0%</td>
<td>6</td>
</tr>
<tr>
<td>Inuit</td>
<td>30</td>
<td>96.7%</td>
<td>0%</td>
<td>0%</td>
<td>3.3%</td>
<td>0%</td>
<td>2</td>
</tr>
<tr>
<td>Itelmen</td>
<td>47</td>
<td>6.4%</td>
<td>0%</td>
<td>14.9%</td>
<td>0%</td>
<td>78.7%</td>
<td>4</td>
</tr>
<tr>
<td>Koryak</td>
<td>155</td>
<td>5.2%</td>
<td>0%</td>
<td>36.1%</td>
<td>1.3%</td>
<td>57.4%</td>
<td>4</td>
</tr>
<tr>
<td>Nganasan</td>
<td>49</td>
<td>2.1%</td>
<td>0%</td>
<td>38.8%</td>
<td>36.7%</td>
<td>22.4%</td>
<td>5</td>
</tr>
<tr>
<td>Ojibwa</td>
<td>28</td>
<td>64.3%</td>
<td>3.6%</td>
<td>7.1%</td>
<td>0%</td>
<td>25.0%</td>
<td>6</td>
</tr>
<tr>
<td>Reindeer</td>
<td>24</td>
<td>37.5%</td>
<td>0%</td>
<td>16.7%</td>
<td>16.7%</td>
<td>291%</td>
<td>5</td>
</tr>
<tr>
<td>Sel'kup</td>
<td>20</td>
<td>0%</td>
<td>0%</td>
<td>35.0%</td>
<td>0%</td>
<td>65.0%</td>
<td>7</td>
</tr>
<tr>
<td>Udehe</td>
<td>45</td>
<td>0%</td>
<td>0%</td>
<td>17.8%</td>
<td>0%</td>
<td>82.8%</td>
<td>5</td>
</tr>
<tr>
<td>Yukagir</td>
<td>27</td>
<td>0%</td>
<td>0%</td>
<td>59.3%</td>
<td>33.3%</td>
<td>7.4%</td>
<td>7</td>
</tr>
</tbody>
</table>

1 = Rubicz et al., 2003; 2 = Lorenz and Smith, 1996; 3 = Merriwether et al., 1995; 4 = Schurr et al., 1990; 5 = Sukernik et al., 1996; 6 = Torroni et al., 1993a; 7 = Torroni et al., 1993b.

Among Native Americans (A, B, C and D), the Aleuts had only two, A and D. Their high frequency of haplogroup D (71.5%) set them apart from Eskimos, Athapaskans, and other Northern Amerindians (see Table 6.2). The Aleuts shared several HVS-1 sequences with other circumarctic populations, but did not have the 16265G mutation which is specifically found among Eskimo groups. They also significantly differ from the Koryak and Even populations of Kamchatka. The Aleuts were closest genetically to the Chukchi and...
MOLECULAR MARKERS IN ANTHROPOLOGICAL GENETIC STUDIES

Siberian Eskimo populations of Chukotka. These results indicate that the Aleuts likely migrated across the Bering Land Bridge and settled the Aleutian Islands from the east, rather than island-hopping from the Kamchatka Peninsula to settle the western Aleutians. This lends further support to the hypothesis that multiple migrations were responsible for the peopling of the New World.

Analytical tools
Understanding the underlying biological and social processes that occur within and between human societies allow anthropological geneticists to make inferences concerning two areas of inquiry that have shaped modern and ancient human populations. The first of these processes focuses on population structure, which may be defined as any number of factors that impact the amount of gene flow or genetic drift within a group or population (Crawford, 1998). The second area of inquiry is population history, which focuses on the extent of biological similarity among groups or populations, reflecting either common ancestry or gene flow (Harpending and Jenkins, 1973). The advent of molecular genetic markers has led to a number of analytical techniques that have either been adopted from classical genetic studies or generated to deal specifically with molecular data. The following is not a comprehensive review of these techniques, but rather offers examples of some of these methods that address questions regarding population structure or history.

Population structure
Several different definitions of population structure exist. A few researchers define population structure as all factors that cause deviation from Hardy-Weinberg equilibrium (Cavalli-Sforza and Bodmer, 1971). Others restrict their definition exclusively to population subdivision, based on cultural or geographic factors (Schull and MacCluer, 1968). The relationship between genetic elements (genes, genotypes, and phenotypes) has also been used to define the structure of human populations (Workman and Jorde, 1980). Population structure can be further subdivided into intrapopulation (variation within a population or subdivisions of a population) and interpopulation (variation between populations) differences. This division may be limited because the boundaries between populations and their subdivisions are often arbitrary and all human populations may be considered a subdivision of a single ancestral population (Crawford, 1998).

Intrapopulation variation
The majority of human populations exhibit some form of internal subdivision due to geographic, linguistic or cultural factors. These subdivisions include language families, clans, tribes, castes, religions, socio-political units or other classifications. Frequently, these aggregates are hierarchical and may serve as effective barriers to gene flow...
between groups, with a varying degree of success. A number of analytical methods are available for evaluating the effects of evolutionary forces on population subdivision. A few techniques that may be useful in quantifying intrapopulation variation and applicable to molecular data are nucleotide diversity statistics, tests of selective neutrality, mismatch distribution, and analysis of molecular variance (AMOVA).

Measures of nucleotide diversity

According to the neutral theory of molecular evolution, the majority of nucleotide substitutions have no effect on fitness of an individual (i.e. are neutral) and most polymorphisms are transient, awaiting fixation due to drift (Kimura, 1968a,b). Therefore, assuming mutation-drift equilibrium it is possible to determine the expected level of diversity (θ) in a population or its subdivision using the mutation rate (μ) and effective population size (N_e) for diploid loci, with the equation:

$$θ = 4N_eμN_eμ$$  \((1)\)

The parameter θ is an important factor in several different molecular statistical techniques and is often compared to the nucleotide diversity measure (π) and the number of nucleotide variant sites, generally shown as θ̄ (S|a_n, where a_n is Σ{a_i} / n) (Waterson, 1975). The π statistic is a measure independent of sample size and analogous to Nei’s gene diversity measure (Nei, 1987). This diversity measure describes the probability that two copies of the same nucleotide drawn at random from the same set of sequences will differ and is represented using the equation:

$$π = n(x_i x_j / n(n - 1))$$  \((2)\)

where n equals the number of sampled sequences, x_i and x_j are the frequencies of i_th and j_th sequences and π_ij is the proportion of nucleotide differences between them.

Measures of selective neutrality

In order to determine whether or not populations are being influenced by evolutionary forces other than selection it is important to ascertain whether the amount of genetic diversity exhibited by these populations deviates from neutrality. Several different neutrality tests exist, including Tajima’s D, HKA, McDonald-Kreitman, Fu’s Fs, Fu and Li’s D as well as others. These statistics have been recently reviewed in detail by Kreitman (2000) and we will only briefly discuss two common neutrality measures (Tajima’s D, Fu’s Fs) used in the anthropological literature. The underlying genetic structure of a population may play a key role in the detection of positive selection (Przeworski, 2002) but factors other than selection such as expansion, bottlenecks, or background selection may also have an impact (Aris-Brosou and Excoffier, 1996; Schneider et al., 2000; Tajima, 1993; Foy and Wu, 1999).
In anthropological genetics, two common measures of selective neutrality, Tajima’s D (Tajima, 1989) and Fu’s Fs (Fu, 1997), are often applied with the assumptions of the neutral theory of evolution. These statistics are appropriate in distinguishing population expansion from constant population size. Population growth generates an excess of mutations in the external branches of the genealogy and therefore an excess of substitutions are present in only one sampled sequence (Ramos-Onsins and Rozas, 2002). This leads to a star-like phylogeny that includes a large central node with several radiating spokes each represented by a single individual.

Tajima’s D (Tajima, 1989) uses information from the sample mutation frequency and is based on the infinite-sites model without recombination. This statistic is appropriate for short DNA sequences or RFLP haplotypes. Tajima’s D compares two estimators of the mutation parameter $\theta$. The test statistic $D$ is estimated as:

$$D = \frac{\theta\pi - \theta s}{\sqrt{\text{Var}(\theta\pi - \theta s)}} \tag{3}$$

where $\theta\pi$ is equivalent to the mean number of pairwise differences between sequences $(\pi)$ and $\theta s$ is based on the number of nucleotide variant sites. Negative scores are indicative of larger values for $\theta s$ relative to $\theta\pi$ signifying the potential effects of population expansion. Positive or statistically non-significant negative scores may indicate the effects of genetic bottlenecks on a population, which tend to create highly fragmented phylogenies and represent inflated values for $\theta\pi$ relative to $\theta s$.

Fu’s Fs (1997) is also based on the infinite-site model without recombination but utilizes data from the haplotype distribution. This test statistic is based on the equation:

$$F_s = \ln\left(\frac{S}{1 - S}\right) \tag{4}$$

where $S$ is the probability of observing a random neutral sample and defined as $S = \text{PR}(k \geq k_{\text{obs}} | \theta = \theta\pi)$, where $(k)$ is equal to the number of alleles similar or smaller than the observed value given $\theta\pi$ and $F_s$ is the logit of $S$. Statistically significant negative scores indicate an excess of alleles, a signature of population expansion. This test is considered less conservative than Tajima’s $D$ and is more sensitive to large population expansions expressed as large negative numbers whereas positive numbers indicate populations impacted by genetic drift (Schneider et al., 2000; Fu, 1997).

Mismatch distribution
Another common method for representing molecular data is through the distribution of pairwise differences, also known as mismatch distribution. This method is applicable to molecular data where differences between alleles can be counted and includes nucleotide substitutions, RFLPs, VNTRs, or STRs (Jobling et al., 2004). The mismatch distribution is constructed by counting the number of differences or scatter plots against the time period.
differences between each pair of subjects and then using histograms or scatter plots to display the frequencies of sites that differ (Rogers et al., 1996). This measure of diversity summarizes the discernible amount of genetic variation within a population. The shape of the mismatch distribution is also highly informative. A unimodal distribution is indicative of population expansion whereas a multimodal distribution indicates constant population size over a long time period (Rogers and Harpending, 1992).

Zlojutro et al. (2006) used mtDNA diversity and a number of these techniques in order to assess intrapopulation variation in Aleutian islanders. Figure 6.4 shows a mismatch distribution for the total Aleut population. This distribution is bimodal with peaks at 0 and at 7, generally interpreted as a signature of long-term population stability. However, others have argued that population substructure and mutation rate heterogeneity may account for multimodal mismatch distributions (Aris-Brosou and Excoffier, 1996; Marjoram and Donnelly, 1994). In order to determine whether or not population structure was influencing the overall mismatch distribution, Zlojutro et al. (2006) separated out three mtDNA haplotypes A3 and D2 (shared by other Native American and Siberian populations), and A7 (characterized by a 16212A transversion) specific to the Aleut population (Figure 6.5). Mismatch distributions for all three haplotypes are unimodal and support evidence for differential population expansion within their respective haplogroups. The A3 haplotype possibly indicates an older demographic event than both the A7 and D2 haplotypes, due to its higher peak at two mutational differences. This latter mismatch distribution is indicative of two population expansions (one for A3 and a later expansion for A7 and D2), which may not have been detected if the underlying mtDNA haplotype structure had not been investigated independently of the total population.

Further support for a dual population expansion in Aleuts is provided through nucleotide diversity scores and neutrality test statistics (Zlojutro et al., 2006). Table 6.3 shows values for $\theta_n$, $\theta_s$, Tajima’s $D$, and Fu’s $F_s$ values within the total Aleut population and
Mismatch distributions for select mtDNA haplotypes (after Zlojutro et al., 2006).

**Table 6.3.** Aleut Diversity and neutrality test statistics for population structure as inferred through mtDNA haplogroups (after Zlojutro et al., 2006).

<table>
<thead>
<tr>
<th>Haplogroup</th>
<th>Theta (SD) A</th>
<th>Theta (SD) B</th>
<th>Tajima's D</th>
<th>Fu's Fs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aleut total</td>
<td>3.338 (1904)</td>
<td>3.582 (1080)</td>
<td>-0.234</td>
<td>-6.678</td>
</tr>
<tr>
<td>Subhaplogroup A3</td>
<td>2.099 (1358)</td>
<td>2.743 (1216)</td>
<td>-0.822</td>
<td>-6.866</td>
</tr>
<tr>
<td>Subhaplogroup A7</td>
<td>0.607 (0.549)</td>
<td>1.223 (0.636)</td>
<td>-1.337</td>
<td>-3.093</td>
</tr>
<tr>
<td>Haplogroup D</td>
<td>0.363 (0.390)</td>
<td>1.085 (0.496)</td>
<td>-1.421*</td>
<td>-4.799*</td>
</tr>
</tbody>
</table>

two mtDNA subhaplogroups (A3, A7) and mtDNA haplogroup D. The Aleuts are characterized by low mtDNA diversity levels with $\theta_A$ equal to 3.338 and $\theta_B$ equal to 3.582. The values are lower for all three of the other haplogroups with the lowest $\theta_A$ (0.363) and $\theta_B$ (1.085) diversity being located in haplogroup D. Fu's Fs statistics are significant for all three haplotypes, while only one, haplogroup D, was found to be statistically significant for Tajima's D. This is indicative of the greater statistical sensitivity of Fu's Fs to Tajima's D, which some researchers consider overly conservative (Ramos-Onsins and Rozas, 2002).

**AMOVA**

AMOVA is an alternate to allele frequency methods and takes into account the molecular relationship of alleles. AMOVA is analogous to a nested analysis of variance (ANOVA) derived from a matrix of squared distances among all pairs of haplotypes. This in turn produces variance estimates and F-statistic analogues designated as $\Phi$-statistics that reflect the correlation of haplotypic diversity at different hierarchical levels of population subdivision. The following equation is used to calculate the total sum of squared deviations (SSD):

$$SSD_{(Total)} = \frac{1}{2N} \sum_{j=1}^{N} \sum_{k=1}^{N} \delta_{jk}^2$$

**Interpretation**

Comparing anthropological genetic diversity and phylogeographic patterns, as inferred from mtDNA sequence analysis, provides valuable evidence for the study of human history and the evolutionary process within and between populations. The highest genetic diversity is observed in specific haplogroups, which can be used to trace the migration patterns and historical events affecting the Aleut population.
where $N$ equals the number of haplotypes, $d_{jk}^2$ is the Euclidean distance between haplotypes $j$ and $k$. This allows for the hierarchical partition of the haplotypes into SSD within populations, SSD within regional groups, and SSD among populations within regional groups. The mean squared deviation (MSD) is obtained by dividing the corresponding SSD by the appropriate degrees of freedom (Excoffier et al., 1992). This method is appropriate for any data where genetic distances between alleles can be calculated.

Rubicz (2001) used AMOVA on mtDNA control region data to determine whether population structure was present between Aleut communities located in western, central, and eastern regions of the Aleutian island chain. Table 6.4 shows the results of an analysis that investigated mother’s place of birth for 47 individuals. This AMOVA indicates significant differences exist between the aforementioned groups and accounted for 27.22% of the variation present in these data. No significant difference was found within groups and the greatest amount of variation was located within communities (70.79%). According to this analysis, Aleuts appear to demonstrate population structure along an east-west axis and these AMOVA results are concordant with both linguistic and archaeological evidence.

### Interpopulation variation

Comparisons between populations have long been of interest to anthropological geneticists due to the information they provide regarding stochastic processes impacting human population structure and phylogeny. These differences may be based on either geographic location or linguistic affiliation of cultural groups and are informative in regards to evolutionary forces affecting genetic variation in these populations. Analytical methods used to study these differences have largely been adopted from classical genetic studies, with a few exceptions such as microsatellite genetic distances and phylogenetic networks. Population relationships are often established through the comparison of genetic distances and displayed either through genetic maps, phylogenetic trees, or in select cases, phylogenetic networks.

Genetic differences and affinities within human populations or subgroups are often measured through allele or haplotype
It is difficult to measure the resulting similarities or diversity between groups by solely viewing the matrices of several alleles or populations. Therefore, a number of genetic distance measures have been developed for comparing variation between populations through the use of summary statistics (Crawford, 1998). A number of genetic distance measures are applicable to molecular data and include those that compute standardized Euclidean squared distances, angular transformations, gene substitutions, and coancestry coefficients (Jorde, 1985). Two commonly used classical genetic distance measures designed for measuring genetic differentiation in subdivided populations are Wright's \( F_{st} \), appropriate for sequence or microsatellite data (Reynolds et al., 1993) and Nei's standard genetic distance, \( D \) (Nei, 1987).

In addition to these techniques, several distance measures have been developed that deal exclusively with microsatellite loci, which follow a stepwise mutational model (Goldstein et al., 1995). Measures applicable to STR or VNTR data include \( \delta p^2 \) (Goldstein et al. 1995), \( R_{ST} \) (Slatkin, 1995), and \( D_{SW} \) (Shriver et al., 1995). However, microsatellite distance measures that rely primarily on mutation (\( \delta p^2 \) and \( R_{ST} \)) are less effective than methods that grant greater weight to genetic drift \( (F_{st} \text{ and } D_{SW}) \) for recognizing population associations (Destro-Bisol et al., 2000; Perez-Lezaun et al., 1997).

Three common techniques for graphically displaying genetic distance data are: genetic maps, phylogenetic trees, or phylogenetic networks. Genetic maps reduce distance matrices into a more manageable two or three-dimensional graphical representation based on various matrix algebra methods such as principal coordinate analysis (Jorde, 1985). This method is mathematically sound and distortions common to phylogenetic tree building are often erased when using gene maps. Phylogenetic trees not only contain information about the relationships between populations but also provide data regarding the fissioning of groups and the time of divergence (Crawford, 1998). Several different methods of tree construction utilizing different assumptions exist: maximum likelihood, maximum parsimony, neighbour-joining, and unweighted paired group method (UPGMA). Statistical error in tree building can be high and it is recommended that over 30 loci be used along with bootstrapping methods, which may allow for an estimation of confidence limits (Jorde, 1985).

An alternative to phylogenetic trees for certain types of molecular data (mtDNA RFLP, mtDNA control region, and male specific Y-chromosome STRs) are phylogenetic networks. These networks offer an advantage over traditional tree building methods that utilize maximum parsimony or maximum likelihood, because networks can distinguish between irresolvable and resolvable character conflict errors that may occur due to homoplasy. Networks represent 'all most parsimonious trees' by highlighting conflicts in the form of reticulations (equally possible mutation routes between nodes in the network) and interpreted as homoplasy (parallel mutations or

reversals).
reversals), recombination, or sequence errors (Bandelt et al., 1995). The network is sequentially constructed through the addition of consensus points (median vectors) to three mutually close sequences at a time. These median vectors are then inferred as either extinct sequences or extant unsampled sequences within the population. Four different types of networks exist and include: minimum spanning networks (MSN), reduced median networks (RM), median joining networks (MJ), and quasi-median spanning networks (QSN). The underlying assumptions of these networks are that ambiguous states are infrequent and that recombination is absent. These assumptions are met for aforementioned molecular data types (Bandelt et al., 1999).

Figure 6.6 illustrates a generalized phylogenetic network of worldwide human mtDNA haplogroups and macrohaplogroups along with their characteristic RFLP, HVS-I and HVS-II sites. Macrohaplogroups consist of a number of haplogroups shared by several populations. A number of these haplogroups are considered to be continent-specific and integration between geographic regions is rare. Sub-Saharan African populations are all characterized by the L-macrohaplogroups (L1, L2, and L3). The macrohaplogroup L3 is thought to have diverged into the macrohaplogroups M and N, which arose in northeastern Africa and subsequently spread throughout the rest of the world. The final macrohaplogroup R then diverged from N. European populations belong almost exclusively to N and R as the haplogroups
H, I, J, N, T, U, V, W and X make up 98% of the total mtDNA variation (Mishmar et al., 2003). Asian populations consist of all three non-African mtDNA macrohaplogroups (M, N and R) and haplogroups (A, B, C, D, E, F, G, X, Y and Z), and, as previously mentioned, Native Americans are almost exclusively composed of haplogroups (A, B, C, D and X).

Genes and language
A research area that has long been of interest to anthropological geneticists is the relationship between linguistic affiliation and genetics (Cavalli-Sforza et al., 1994). Languages are often considered to be potential barriers to reproductive success of a population and may influence the population structure found within a region. The nature of the language–gene relationship is often controversial and the results are frequently dependent on both the linguistic classification and analytical method utilized. Some areas of the world demonstrate a clear relationship between genetic and linguistic distances, while for other regions the association is more ambiguous or absent (Crawford, 1998). One example of a research study that examined this relationship between American Na-Dene and Central Siberian Yeniseian speakers was conducted by Rubicz et al. (2002) in order to test Ruhlen’s (1998) hypothesis that these two linguistic families were closely related. This relationship was considered controversial because of the large geographic distance separating the speakers of these two language families.

Rubicz et al. (2002) investigated this genetic-linguistic relationship with several Native American and Siberian populations using blood group polymorphisms and four mtDNA haplogroups (A, B, C and D). A common genetic map technique known as R-matrix analysis was utilized in order to visualize the association between the two regions. R-matrix is a relational statistical technique that allows for the representation of population structure in two or three dimensional space. The first step is to calculate a genetic distance matrix using the equation:

\[ R_{ij} = (p_i - \bar{p})(p_j - \bar{p})/(\bar{p}(1 - \bar{p})) \]  

(6)

where \( R_{ij} \) is the kinship coefficient for every allele, \( p_i \) and \( p_j \) are the allele frequencies in populations \( i \) and \( j \) and \( \bar{p} \) is the weighted gene mean frequency of allele \( p \) in the matrix. The genetic distances are then averaged into a variance-covariance R-matrix. The R-matrix is reduced into eigenvectors that correspond to the percentage of variation observed in the matrix and plotted using principal component analysis (Harpending and Jenkins, 1973).

Figure 6.7 shows the plot of the first two principal components of an R-matrix using 17 populations and the four mtDNA haplogroups (A, B, C and D) from Rubicz et al. (2002). The first two eigenvectors account for 78.5% of the variation found within the sample populations. The first axis separates Native American groups from Siberian populations, with the exception of Asian Eskimo who group with
northern Native Americans. The second axis divides northern and southern Na-Dene speakers. Figure 6.8 shows the plot of the mtDNA haplogroup and illustrates that the single extant Yeniseian speaker population (Kets) are separated from Na-Dene speakers (Haida, Dogrib, Apache, and Navajo) by the presence of mtDNA haplogroup C and the absence of haplogroup A. Along with evidence provided by additional analytical techniques, Rubicz et al. (2002) concluded that the available genetic data demonstrated that Na-Dene speaking populations clustered with other Native American groups, while the Kets resemble surrounding Siberian populations and that spatial patterning accounts for the majority of variation present in these populations.

Genes and geography
The relationship between genetics and geographic subdivision is a second area that allows anthropological geneticists to generate
inferences regarding interpopulation structure. This relationship reflects the 'isolation-by-distance model' and is based on the assumption that populations proximal geographically will demonstrate a higher genetic affinity than those groups found to be at greater spatial distances. Crawford et al. (2002) investigated the relationship of two Siberian populations, the Evenki from central Siberia and the Kizhi of Gorno Altai from southern Siberia, to surrounding populations using a number of classic, coding and non-coding molecular markers including ADH, ALDH, COL1A2, mtDNA, Y-chromosome, VNTRs and STRs. Figure 6.9 shows a phylogenetic tree of four autosomal STR loci (TPOX, CSF1PO, THO1, and vWA) among three Siberian populations (Ket, Altai, Evenki) along with four other populations (Asian, Caucasian, Javan, and Amerinds). The genetic distances for this tree were calculated using the Shriver’s $D_{SW}$ distance (Shriver et al., 1995) and a dendogram was constructed utilizing the Fitch-Margoliash method (Fitch and Margoliash, 1967). The first bifurcation of the tree separates the Evenki from the other sample populations. The Altai separate out next and are distinct from the other Siberians, Asians, and Amerinds. The Amerinds and Caucasians cluster together as do the Asians and Javans. Based on this evidence, along with results from other markers, it was concluded that the Evenki have maintained their genetic uniqueness as opposed to the Altai who demonstrate a closer relationship to European populations.

A problem for genetic distance measures and their association with other distance matrices (such as geography) is their inability to provide measures of statistical significance to the relationships demonstrated between populations. A commonly used method that does allow for comparisons between matrices and testing of statistical significance is the Mantel test (Mantel, 1967). This method tests

![Phylogenetic tree based on four STRs (TPOX, CSF1PO, THO1, and vWA) for three Siberian population (Ket, Altai, Evenki), Asian, Native American and European population. Genetic distances calculated using Shriver's method (Shriver et al., 1995) and tree was constructed using the Fitch-Margoliash method (Fitch-Margoliash, 1967).](image)

Adapted from Crawford et al., 2002.
the association between the elements of two matrices using the equation:

\[ Z_{AB} = A_{ij}B_{ij} \]  

(7)

where \( A_{ij} \) and \( B_{ij} \) are elements of row \( i \) and column \( j \) of matrices \( A \) and \( B \), resulting in an unnormalized correlation coefficient. The statistical significance of the correlation is tested by comparing the observed correlations against a sample distribution of \( Z \) based on randomized \( B \) matrix (Crawford and Duggirala, 1992).

Other factors besides geographic distance may have an impact on the relationship between genetics and geography. These factors include past geological and ecological events that may have isolated human populations and contributed to present day genetic makeup of these groups. Crawford et al. (1997) used Mantel’s test to investigate potential ecological events that occurred during the Pleistocene and early Holocene in Siberia and may have contributed to the genetic isolation of these groups. Based on Mantel’s test results they found that 30.6% of genetic variation in the region could be explained through the joint effects of geography and language. However, partial correlations between all three matrices (while holding geography constant) demonstrated that the relationship between genetics and linguistics was statistically insignificant. Crawford et al. (1997) concluded that most of the genetic differentiation in Siberia is due to geographic patterning. This may be due to the location of Pleistocene glaciers of northern Siberia as well as the presence of Lake Mansi, which played active roles in the geographic isolation and linguistic differentiation in the region.

Population history

Population history focuses on the genetic resemblance of populations, reflected through common descent or gene flow (Harpending and Jenkins, 1973). These two aspects of genetic similarity are normally impossible to differentiate and are closely related to interpopulation variation. However, population history is concerned with past evolutionary events and their timing, which may have impacted the current genetic makeup of populations, whereas interpopulation variation may reflect the existing biological relationships between living populations. Molecular markers and the advent of coalescent theory (Hudson, 1990; Kingman, 1982) have enhanced the ability to determine both the effects of stochastic evolutionary events that shaped a population and the relative time that they occurred. A number of coalescent dating techniques have been developed and aid in providing a context for relating genetic data with other forms of prehistoric data. Three dating methods that may be applicable to anthropological genetic research are mismatch distributions (Rogers and Harpending, 1992), p (Forster et al., 1996; Saillard et al., 2000), and the averaged square distance (ASD) between a root microsatellite haplotype and all other haplotypes (Goldstein et al., 1995).
Techniques for estimating population history from sequence data are frequently based on a group of mathematical models collectively referred to as coalescent theory (Kingman, 1982). This theory is useful for the characterization of statistical properties located within intrapopulation phylogenies and is used to estimate effective population size, recombination, and migration rates from sampled sequence data. The mutual history of the sampled sequences creates a genealogy, the lineages of the sample sequences converge or 'coalesce' backwards in time, until the most recent common ancestor of the sample is reached (Pybus and Rambaut, 2002). The coalescent process is an approximation of classic population genetic models and is valid when effective population size is large (Strimmer and Pybus, 2001). Coalescent theory has several useful applications for anthropological geneticists including mathematical modelling, simulation tool for hypothesis testing, inferential tool, and exploratory data analysis (Rosenberg and Nordborg, 2002).

**Chronometric techniques**

In order to make sense of population history, anthropological geneticists are challenged to place it within a broader context, which is most readily achieved through chronometric techniques. The dating of molecular polymorphisms is dependent on the theory of the molecular clock. According to this theory, genetic variation (consisting of mutation, recombination, and genetic drift) accumulates at a predictable rate (Jobling et al., 2004). The measurement of this rate can be thought of as the speed at which the clock ticks. This can be done directly, through the observation of mutations occurring in rapidly mutating markers such as microsatellites, or indirectly by comparison of closely related species in conjunction with their estimated time of divergence (based on the fossil record). Typically this method assumes that natural selection has not been operating on the loci under investigation. Non-recombining portions of the genome, the MSY and mtDNA, have been extensively used for dating because in these cases mutation alone is assumed to drive the molecular clock. The number of mutations between lineages can be directly counted and related to the time of the most recent common ancestor, without the confounding effects of recombination. After estimating the age of a particular node, it may be related to a particular event (such as the split between two human populations) in association with a larger anthropological question such as the timing of the peopling of the New World by migrations originating in Asia (see Salzano, Chapter 15). In this way, molecular data can be used to establish a chronological record of events in human prehistory. Three molecular dating techniques common in the anthropological literature are mismatch analysis, $\rho$, and ASD.

Mismatch analysis is also a useful statistical approach for estimating time of expansion for recently established populations. Rogers and Harpending (1992) demonstrated that pairwise differences
FORCES OF EVOLUTION

The modern synthesis of evolutionary theory indicates that there are four forces of evolution, all resulting from deviations of the underlying assumptions of Hardy-Weinberg-Castle genetic equilibrium.

between nucleotide sequences increased by a rate of $2\mu$ (where $\mu$ equals the mutation rate) for each generation during population growth. From this estimated substitution rate, it is possible to estimate $N$ (population size) of a sample prior to population expansion. It is also possible to use coalescent data (Hudson, 1990) to estimate the initial timing of population growth in mutational units using the equation:

$$\tau = 2\mu t$$

where $t$ is time in generations and $\mu$ is the mutation rate. Taking the parameters $\theta_0$ and $\theta_1$ as the population estimates before and after expansion, respectively, and fitting these and $\tau$ through the least squares method to the observed mismatch distribution permits an estimate of expansion time in mutational units over divisions of time (Rogers and Harpending, 1992). Using the mutation rate of the molecular marker as an estimate, an absolute date can be calculated and then be used to test hypotheses about the history of a population by comparing data from other sources such as archaeology or ecology.

Some researchers have cautioned that underlying intrapopulation structure may have a profound impact on the chronometric dates provided by mismatch analysis, and that phylogenetic dating should take into account the root phylogenetic structure of the population (Zlojutro et al., 2006).

Two commonly used chronometric techniques that require a specified root haplotype are $\rho$ (Forster et al., 1996; Saillard et al., 2000) and ASD (Goldstein et al., 1995). Both statistics are expressed as being equal to $\mu$ (mutation rate) multiplied by $t$ (time in generations), however the methods are appropriate for different categories of molecular markers. ASD is restricted to microsatellite haplotypes, but $\rho$ is applicable to any marker that produces haplotype data. The $\rho$ statistic is an intra-allelic diversity measure that requires the construction of a phylogeny, most often a network, and is representative of the average number of mutations between the root haplotype and every individual in the sample. These mutations are quantified from the network itself in order to account for mutation rate heterogeneity or homoplasy occurring with the sample data. A phylogenetic reconstruction is not necessary for ASD as the equation that it is derived from corrects for homoplasy and mutation rate heterogeneity (Goldstein et al., 1995). Different methods of chronometric dating often result in temporal dates that are not always concordant with each other and some caution should be exercised in extrapolating inferences solely from molecular data.
These forces of evolution change the frequencies of alleles in a gene pool, over time. The concept of genetic equilibrium, as first verbalized by Castle and shortly thereafter by Hardy and Weinberg, assumes: (1) populations of infinite size; (2) panmixis (random mating); (3) equal genetic contribution of each genotype to the next generation. Given these conditions, the frequencies of genes remain constant from generation to generation. Changes in the frequencies of genes over time constitute evolution.

A discussion of the forces of evolution played a prominent role in the 1973 volume on Methods and Theories of Anthropological Genetics with at least four chapters covering various facets of selection, gene flow, and genetic drift. Although mutations were not specifically covered by a chapter, the generation of variation as observed in genetic markers was discussed in one chapter. However, in this volume on Anthropological Genetics, the remainder of this chapter focuses on all four of the forces of evolution and how the molecular revolution has affected their study and measurement.

Natural selection
Darwin’s major contribution to science was the development of the theory of evolution, guided through the actions of natural selection. He stressed survival of the fittest and the action of mortality as the driving force of evolution. However, he did not understand either the sources of normal genetic variation or the mechanisms of heredity. Ronald Fisher in 1930 helped create a new synthesis in evolutionary theory with the publication of his tome ‘Genetical Theory of Natural Selection’. Like Darwin, Fisher emphasized natural selection as the engine of evolution.

Selection operates entirely through differential mortality and fertility. It is not merely the survival of the individual but the successful reproduction and the transmission of the genes. In a given environment the individuals who are the fittest, survive, reproduce and make up the subsequent generations of the population.

Until recently there were few documented cases of natural selection detected in human populations. Prior to the availability of molecular data, in order to demonstrate the action of natural selection in human populations significant differences in fertility had to be ascertained when comparing specific genotypes or genes. Since fitness differences between genotypes in a given generation were often minute, except for rare deleterious or fatal mutations, studies of selection required exceptionally large sample sizes. The best two examples of selection operating on humans involve genes associated with resistance to malaria and with birth weight.

Selection and extremes in birth weight
Karn and Penrose (1951), based upon 6,693 English female babies, showed that selection operates against both the very small
gene (particularly premature) and very large births. They demonstrated that only 41% of infants born weighing less than 4.5 lbs survived 28 days after birth. The additional mortality due to selection acting on birth weight was computed by subtracting the fitness of the optimal set of phenotypes ($S_0 = 0.985$) from the overall fitness ($S = 0.959$) which equals 0.026 (Spuhler, 1973). However, during the last five decades with the technological developments associated with treatment and maintenance of premature births, the survivorship of small infants has increased exponentially.

**Adaptation to malaria**

The best example of selection operating on human populations is the association between malaria and a series of mutations that alter the environment of the erythrocyte and interfere with the life cycle of the malarial parasite. According to World Health Organization, approximately 1 to 2 million people (mainly children) die from malaria each year. A disease with such an ancient evolutionary history, high mortality and lowered fertility is likely to be acted upon by natural selection. A number of hemoglobinopathies (hemoglobin $S$, $C$, $D$, $E$, $\alpha$- and $\beta$-thalassemia), the absence of a chemokine receptor (Duffy null), and low levels of the RBC enzyme glucose-6-phosphate dehydrogenase ($G$-6-$P$D deficiency) affect the life cycle of the *Plasmodium* organism and increase the fitness of the carriers in malarial environments. However, the specific mutations that provide resistance to malaria vary by geographical region where the chance mutations (particularly in the $\beta$-globin region of the genome) had occurred. For example, a suite of mutations in Africa include: $S$ (sickle cell anemia), $C$ hemoglobin, $\alpha$-thalassemia, $G$-$6$-$P$D deficiency, and Duffy null blood group. While in the Mediterranean region, $\beta$-thalassemia, $G$-$6$-$P$D deficiency and hemoglobin D provide resistance to malaria. The dramatic spread of hemoglobin $S$ mutation in Africa accompanied the introduction of slash-and-burn agriculture from Malaysia and demonstrated how culture and genetics interact in human populations.

**Applications of molecular markers to selection**

Molecular genetics offers a new dimension to the study of selection: with the sequencing of the coding regions of the genome it is possible to detect the signature and possibly the type of selection. By comparing sequences in a coding region of the genome to a region that is neutral selectively it is possible to determine whether balancing, directional selection or no selection may have been involved. In balancing selection, the region shows a higher level of sequence diversity, an excess of intermediate-frequency variants and a positive
value for Tajima's D (Bamshad and Wooding, 2003). With directional selection a lower level of sequence diversity, an excess of low-frequency variants are coupled with a negative value of Tajima's D. A beneficial gene spreads rapidly and widely reducing genetic variation. Table 6.5 summarizes from the recent literature regions of the genome that selection may have acted upon and the type of possible selection.

Table 6.5 summarizes the signatures of natural selection in four regions of the genome that have been associated with factors that may affect fertility and mortality in human populations. As expected, the β-globin region and G-6-PD deficiency involve the signature of selection, but through different forms of adaptation. The β-globin region contains mutations for hemoglobin S, which is maintained as a balanced polymorphism with the heterozygotes having a selective advantage over the homozygotes. On the other hand, directional selection appears to be operating in favour of lower levels of G-6-PD, with fewer deleterious effects in homozygous females or hemizygous males. However, individuals who are designated as genetically deficient for G-6-PD, still synthesize some enzyme, although at a lower level. A recent review article by Kwiatkowski (2005) examines the relationship between common erythrocytic variants that affect resistance to malaria. CCR5 gene encodes a cell surface receptor for mediator molecules (chemokines) for HIV-1 virus and bears the signature of selection (Bamshad et al., 2002). The sequence variation in the CCR5 region is substantially higher than expected and deviates significantly from tests of neutrality. One mutation in the CCR5 region, Δ32, has been linked with bubonic plague and prevents both HIV and possibly Yersina pestis from entering human macrophages. However, a recent study has questioned whether bubonic plague could extract the intensity of selection necessary to account for the observed signature and suggests that smallpox is a more likely agent of selection at the CCR5 region (Galvani and Slatkin, 2003).

Mishmar et al. (2003) have recently presented an argument that the colonization of Central Asia and Siberia by humans from Africa resulted in a five-fold expansion of mtDNA diversity. They also found that the Siberian and Native American mtDNA haplogroups, A, C, D, G, Z, Y and X showed significant deviation from neutrality for Fu and Li (D*) test but not Tajima's D. Under conditions of extreme cold, Mist humans by mtDNA require a dietary ATP. Although population variants of intake, or of these variants available, between humans can be fruitfully.

Mutation
Mutation is which occur through the material. The radiation, energy level, germ cells in order for transformation mutation is which all other.

The Dutch suggest the experiment (Oenothera lamarckiana) through simple onions elonged by Re evening prim plants were hypothesis led to mutation in breeding experiments mutations of concluded that genes by more than natural selection unfit genotype.

Morgan’s in the formation of Fisher (1930) individuals perform genetic drift, much more.

In the 1960s,
cold, Mishmar et al. have argued that selection had operated on humans by favouring the increase of heat production. Variants of mtDNA reduce ATP production but increase heat production and require a higher caloric intake provided by a high-fat diet. This dietary-ATP link would provide an advantage under cold stress. Although higher basal metabolic rates are observed in arctic populations it is unclear whether the elevated metabolism is due to variants of mtDNA and oxidative phosphorylation, higher caloric intake, or increased levels of thyroxine, or the interaction of all of these variables. With the molecular genetic tools that are now available, such hypotheses can be tested and the relationship between human molecular characteristics and physiological function can be fruitfully explored.

Mutation

Mutation is the random introduction of new alleles into a population which occurs when the DNA structure of a chromosome is changed through the insertion, replacement, deletion, or inversion of genetic material. These genetic changes may be caused by factors such as radiation, extreme temperatures, or chemical mutagens. On a species level, germ cell mutation is the source of all new genetic material and in order for changes to be passed from one generation to the next, transformations must occur in the gametes' DNA. Therefore, mutation is the creative basis of evolution and is the foundation on which all other forces operate.

The Dutch botanist, Hugo De Vries (1901–1903), was the first to suggest the importance of mutation in evolutionary theory. Through the experimental investigation of the evening primrose plant (Oenothera lamarkiana), De Vries suggested that speciation occurred through single mutations. This hypothesis was subsequently challenged by Renner (1914) and Cleland (1923) who demonstrated that evening primrose is a permanent heterozygote and that De Vries' plants were segregates of this rare genetic form. However, this hypothesis led Morgan (1925, 1932) to propose a more gradual role of mutation in evolution. Morgan et al. (1915) conducted extensive breeding experiments on Drosophila melanogaster and showed that mutations occur in small but quantifiable frequencies. Morgan concluded that evolution occurs through the substitution of existing genes by more beneficial mutations in a population and that the role of natural selection was to preserve useful mutations and remove unfit genotypes.

Morgan's mutation-selection theory began to lose popularity with the formation of the synthetic theory of evolution as proposed by Fisher (1930), Wright (1931), and Haldane (1932). These three individuals performed extensive mathematical formulations on mutation, genetic drift, and natural selection and concluded that selection was much more effective in changing gene frequencies than mutation. In the 1960s the synthetic theory was adopted by the majority of
biological scientists who characterized mutation as: (1) being random and recurring at a reasonably high frequency; (2) the principal source of variation; (3) natural populations contain a large amount of genetic variability due to previous mutations; (4) due to this high degree of genetic variability, populations have a certain degree of evolutionary flexibility and do not have to evolve in response to changes in the environment; (5) the majority of advantageous mutations should remain fixed due to the high mutation rate; and (6) evolution of new species occurs gradually through the accumulation of advantageous mutations that are acted on by natural selection (Nei, 1987).

In the late 1960s, Kimura (1968a,b) and King and Jukes (1969) began to question some of the basic assumptions of the synthetic theory of evolution through the experimental investigation of newly discovered techniques for the identification of protein polymorphisms. Kimura (1968a, 1983) investigated amino acid substitution rates and allozyme polymorphisms in a number of different species and argued that their genetic load would be too great if explained only by selection. He concluded that these protein polymorphisms were more easily explained through the interactions of genetic drift and neutral mutation rather than solely by natural selection. These findings led researchers to propose the neutral theory of evolution based on the principle that nearly all mutations are selectively neutral and that the majority of molecular variation is driven through genetic drift rather than natural selection. The basic tenet of this theory is that through genetic drift, new neutral alleles become more frequent within a population. Over time these neutral mutations decline or disappear, or in rare cases become fixed within a population or species. When fixation of an allele occurs they accumulate within a previously fixed sequence of the genome. The advent of DNA sequencing technology in the 1980s provided further support for the neutral theory and it now forms the foundation of molecular DNA analysis. Neutral theory has led to a number of advances in molecular biology, including the ability to test null hypotheses using neutrality test statistics and the formation of the molecular clock hypothesis, both of which have been previously discussed in this chapter.

Applications of molecular markers to mutation

Mutations form the foundation of molecular anthropology as both sequence differences and protein polymorphisms are the starting points for drawing inferences regarding the structure or history of a population. The ability of molecular technology to resolve single mutations within the entire human genome has greatly enhanced the ability to formulate hypotheses about the other evolutionary forces impacting human populations. Molecular technology also allows for the identification of ancestors, such as the radiation of Homo sapiens, and the determination of the geographical origins of populations.
APPLICATIONS OF MOLECULAR MARKERS TO MUTATION

Using random amplified polymorphic DNA allows for the study of the effects of different causes of mutation such as those genetic differences caused by long-term exposure to radiation.

Dubrova et al. (2002) investigated germ-line mutations at eight autosomal human minisatellite loci (D2S90, D1S172, D10S180, D10S473, D1S7, D7S21, D1S8 and B6.7 located on chromosome 20q13) in rural areas of the Ukraine and Belarus that were heavily contaminated by radiation from the Chernobyl nuclear accident. Using control and exposed groups composed of families with children conceived before and after the Chernobyl accident they found a significant 1.6-fold increase in the germ-line mutation rate of exposed fathers. They did not find an elevated mutation rate in exposed mothers. They suggested that this resulting sexual dimorphism may be due to differences in the timing of spermatogenesis and oogenesis. Spermatogenesis is a continuous process of mitotic and meiotic cell divisions occurring in males at the onset of puberty, whereas female oocytes are formed in the late stages of embryogenesis and remain inactive until puberty. All females used in this study were at least eight years old at the time of the Chernobyl accident, indicating that ionizing radiation may be unable to affect the stability of minisatellite loci in the maternal germline. Based on these results they concluded that for the males of this population the elevated mutation rates of minisatellites were caused by post-Chernobyl radioactive exposure. They also indicated the need for future research on the mechanisms of radiation-induced mutation at human minisatellites.

Gene flow

This force of evolution has been defined by Workman (1973: 117) as 'the intermixture among individuals from different populations'. It can also be viewed as the addition of new genes into a gene pool. Migration rates and hybridization determine the degree of gene flow into and out of specific populations. Gene flow increases genetic variability within populations (i.e. increases heterozygosity) and decreases genetic differences among populations. Gene flow makes it possible for mutations arising in one population to be exposed to a new environment with its unique selective factors.

Estimates of admixture in human populations can be traced to the original formulation of Bernstein (1931) which was based on the proportionate contribution of one population to the genetic makeup of the hybrid, expressed as:

\[ m = q_a - Q/q - Q, \]  \hspace{1cm} (9)

where \( m \) is the proportion of admixture, \( Q \) and \( q \) are the allelic frequencies in the two parental populations, and \( q_a \) is the frequency of the same allele in the hybrid population. This method was first applied to Brazilian populations (Ottensooser, 1944; DaSilva, 1949) and then to Afro-American US populations by Glass and Li (1953). Since the use of single loci gave vastly different estimates of \( m \),
Roberts and Hiorns (1962), Pollitzer (1964) and Elston (1971) used multiple alleles and various multivariate methods to estimate \( m \) in tri-racial populations (see Crawford et al., 1976 for a discussion of multivariate approaches, such as maximum likelihood, true least squares, multiple regression to the measure of admixture).

**Application of molecular data to the measurement of admixture**

Molecular markers offer greater precision than classic markers (such as blood groups and protein) for the measurement of admixture because they are more informative and maternal migration (using mtDNA) can be distinguished from paternal (male) migration (using MSY). However, autosomal microsatellites have been effectively utilized to measure admixture in various populations. For example, Destro-Bisol et al. (1999), using ten unlinked microsatellites plus a linked parital \( \text{Alu} \) deletion, estimated admixture by means of a maximum likelihood procedure in an African American population. The percentage of European admixture was 26\% \pm 2\%, a narrow confidence interval indicating that the allele frequency data from the microsatellite frequencies for the hybrid and parental populations provided a reliable estimate.

An elegant study by DiBenedetto et al. (2001) applying autosomal markers, MSY and mtDNA, permitted the testing of specific historically derived models concerning the nature of the gene flow from Central Asia into the Anatolian peninsula approximately 1,000 years ago. They tested three models: (1) elite dominance — i.e. a small portion of an invading army imposes the Turkic language on the inhabitants of Anatolia but do not contribute genetically to the Anatolian gene pool. This model was contradicted by the substantial Central Asian contribution, with a median rate of admixture of 31\% based on Y-chromosome markers; (2) massive instantaneous population movement in the form of a military invasion. This model predicts greater incidence of Central Asian haplotypes at Y-chromosome markers than mtDNA. This prediction was not confirmed; (3) small numbers of Oghuz Turks invaded Anatolia, followed by continuous gene flow from Central Asia, at about 1\% for 40 generations. The introduction of the Turkic language into the population of Anatolia facilitated long-term, continuous gene flow. This model best fits all of the available data.

**Genetic drift**

Genetic drift is the random fluctuation of gene frequencies from one generation to the next, resulting in unpredictable evolutionary outcomes. It is the accidental loss (or fixation) of alleles due to sampling error, the effects of which are most pronounced in small populations. Within a population, genetic drift acts to reduce heterozygosity and causes the loss of genetic variation between individuals.

Genetic drift is a stochastic process that can affect the frequency of alleles in a population over time. It can lead to the fixation of alleles at a higher or lower frequency than expected under neutral conditions. In the absence of selective pressures, drift can cause the random loss of rare alleles from a population or the fixation of rare, deleterious alleles.

In the absence of drift, the rate of fixation of a new mutation in a population is given by

\[
N_e \mu = \frac{1}{2} \left( 1 - \frac{1}{2N_e} \right)
\]

where \( N_e \) is the effective population size and \( \mu \) is the mutation rate. This formula shows that the rate of fixation decreases as the effective population size increases, reflecting the greater opportunity for drift to occur in smaller populations.

In summary, genetic drift plays a crucial role in shaping the genetic diversity of populations. Understanding the forces that drive drift helps us to better comprehend the evolutionary history of species and the genetic makeup of contemporary populations.
causes the loss of alleles. The long-term effects of genetic drift are the reduction of variation within a population and the increase of variation between populations, leading to their divergence.

Gulick was the first to note the importance of random evolutionary processes in small isolated populations (Gulick 1872); however, it was Sewall Wright who developed the theory of genetic drift. In the 1930s and 1940s he worked out a series of mathematical models to deal with the effects of random processes in populations of finite size (detailed in Wright, 1969). The characteristics of random processes can be summarized as follows: (1) it is not possible to predict the direction of change; (2) the magnitude of change depends on the size of the population, with smaller populations experiencing larger changes; and (3) two samples taken from a population will differ not only from the parental population, but also from each other. According to Wright, the repeated random sampling of a population may cause genetic drift to occur, and this can have an evolutionary impact on a species.

Genetic drift can cause the loss of alleles from a population and reduce its heterozygosity (Wright, 1931). Alleles that are initially present in small frequencies in a population are more likely to be lost, while those present at higher frequencies have a greater chance of going to fixation. Most new mutations that arise in a population are transitory. However, if the population is small and isolated, a new mutation has an increased likelihood of becoming fixed through genetic drift. Fewer generations are needed for the fixation or elimination of alleles than in large, panmixtic populations. In addition, individuals are more likely to be identical by descent in small populations, meaning they share certain ancestry. This contributes to the reduction of heterozygosity, and can result in a decrease in fitness. Recessive alleles that are normally rare in other populations will have a greater chance of being homozygous and expressing genetic diseases.

Unique historical events can have an impact on human populations, making them more susceptible to the actions of genetic drift. Two examples of situations that cause a reduction in population size include population bottleneck, in which there is decrease in size of an existing population, and founder effect, where a subset of the original population establishes a new colony. These events are themselves categorized as types of genetic drift.

In 1931, Wright introduced effective population size as a means by which the magnitude of genetic drift can be measured (Wright, 1931). The relationship between effective population size ($N_e$) and the census size of the population ($N$) is expressed as:

$$N_e = N + 1/2.$$  

(10)

In the absence of natural selection and mutation, $N_e$ allows for the calculation of the probability that a new allele will become fixed, and the rate at which this may occur. The smaller the effective population size, the greater the effect drift will have on it. $N_e$ is nearly always
smaller than \( N \), and is difficult to measure because it is affected by a number of parameters, including an uneven number of males and females, and variation in family size. Fluctuations in the overall size of a population will also affect its evolution, and reductions will be disproportionately important because \( N_e \) is shown to approximate the harmonic mean of the population (Wright, 1938; Crow and Kimura, 1970). Most large populations are subdivided into smaller groups (demes) that are partially isolated from each other, and whose members are more closely related to each other than to the population as a whole. Thus, population structure also acts to reduce \( N_e \).

In the 1950s and 60s there was a surge in studies of reproductively isolated, inbred human populations in order to better understand the effects of genetic drift and transmission of rare genetic diseases. Examples include studies of the populations of Tristan da Cunha, Parma Valley, St Bartholemew, the Habbanite Jews, and various Anabaptist groups such as the Amish, Dunkers, and Hutterites (Goldschmidt, 1963). Among the Amish, McKusick et al. (1964, 1971) described four rare recessive genetic disorders, each of which was present at a relatively high frequency in a different community: Ellis-van Creveld syndrome in Lancaster County (Pa.) Amish; Pyruvate Kinase deficient hemolytic anemia in Mifflin County (Pa.) Amish; Hemophilia B in Holmes County (Ohio) Amish; and limb-girdle muscular dystrophy in Amish of Adams and Allen Counties, Indiana.

Application of molecular data to the measurement of genetic drift

Today, molecular markers facilitate investigations into the effect of random genetic drift on human populations because allele frequencies can be assessed directly rather than relying on observations of phenotypic changes, such as rare recessive disorders, or the quantification of gene products.

One example of the application of molecular data to assess the effects of genetic drift in a population is the use of mtDNA and MSY markers in investigating the population history of the Tristan da Cunha islanders by Soodyall et al. (1997, 2003). Tristan da Cunha is a small volcanic island in the southern Atlantic Ocean, discovered by the Portuguese in 1506. In 1816 the British placed a garrison on the island to prevent the French from attempting to rescue Napoleon Bonaparte from exile on St Helena, 1,300 miles to the North. The British garrison was removed a year later, at which point Corporal Glass, along with his wife, their two children, and two other men, stayed behind to establish the first permanent residence on Tristan. Genealogical records, going back to those first settlers, have been carefully reconstructed for all the island's inhabitants. The current population of 278 is described as having descended from 15 known individuals, all of western European ancestry. There is also historical
is affected by the number of males in the overall reductions to approx. 38%: Crow and I into smaller herds, and whose population size is reduced. Nonreproductive the overall understanding of the diseases. da Cunha, and various Hutterites (1964, 1971) of which was immunity. Ellis, Trisch: Pyruvate [Pa] Amish; and limb-girdle muscles, Indiana.

**Measurement**

To assess the effect of allele frequency observations, the Tristan da Cunha is a discovered byarrison on the cave Napoleon the North. The point Corporal to other men, once on Tristan. ers, have been. The current from 15 known also historical documentation of two population bottlenecks which occurred on Tristan, one in the 1850s when the original founder died and 25 of his descendants left, followed by the emigration of a pastor and 45 additional members of the community (Roberts, 1968). The second occurred in 1885 due to a boat capsize that killed 15 adult males, leaving only four on the island (one of which was mentally retarded and two were elderly). In 1997, Soodyall et al. compared the genealogical information with mtDNA data, and found both lines of evidence indicated the current Tristan population was derived from only five females. Documentation of two sister pairs among the original female founders was not supported by the mtDNA data, however, which only found evidence for one pair. Y-chromosome analysis of Tristan males demonstrated that while seven surnames were in use in the living population, more than seven males had contributed to the genetic makeup of the living population (Soodyall et al., 2003). Two additional Y-chromosome types were present that did not match with the recorded surnames, one of which could be accounted for by a single-step microsatellite mutation from a legitimate Y haplotype, while the other was not indicated in the records and was of obvious non-island origin. In these studies, when molecular markers were compared with the written records of Tristan da Cunha, they pointed out errors in the genealogical information, and provided additional documentation of the action of founder effect and bottlenecks in this small island population.

In another study, molecular markers were used to compare the genetic makeup of the Aleuts of Bering Island, Russia with that of their founding population in the Aleutian Islands and to determine the genetic consequences of unique historical events on this population (Rubicz et al., 2005). Between 1824 and 1828, Aleuts were forcibly relocated from their homeland in the Aleutian Islands to the Commanders (Bering and Medni Islands) to work for the Russian-American Company. Families from Atka, in the central Aleutians, were relocated to Bering, and Medni was settled by Aleuts from Attu, in the western Aleutians. In 1825 only 45 Aleuts resided on Bering, in addition to 15–30 original Russian settlers. By 1826 the number of Aleuts increased to 110. The Commander Aleut population sizes fluctuated over time, with an all-time high of 626 (330 on Bering) in 1892. Depopulation due to epidemics and relocations to the Kamchatka Peninsula reduced this number to 364 (204 on Bering) by 1923. In 1969 the Medni Aleuts were relocated to Bering. Today the Bering population consists of approximately 300 individuals. Rubicz et al. (2005) found that Bering Island Aleut mtDNAs belonged to a single haplogroup, D, whereas the founding Aleut population had both haplogroup D (72%) and A (28%) mtDNAs. While Derbeneva et al. (2002) suggest the fixation of haplogroup D in the Bering population is due to intergenerational drift, Rubicz et al. favour the explanation that it was the result of a founder effect. There is historical documentation that Aleut families, not a random sample of individuals from the parental population, were originally brought to the Commander
Islands. It is more likely that haplogroup A mtDNAs were absent from the Aleut founders of Bering Island, rather than their being lost over the few generations since the establishment of the community, considering they were originally present at a frequency of nearly 30%. This study provides further details of the effect of random processes such as historic events on small isolated communities.

Conclusion
This chapter demonstrates the essential nature of molecular markers to the study of human evolution and the out-of-Africa diaspora, topics central to the field of anthropological genetics. These markers, mtDNA, STRs, VNTRs, SNPs, MSY and the sequences of the coding regions of the genome, are much more informative than our early markers, such as blood groups and proteins. The molecular markers characterize the gene pool (total genetic makeup) of a population and bear witness to the indelible signatures of the forces of evolution. The molecular markers have provided researchers with a unique set of tools that can be used to reconstruct sex specific population history and a molecular chronology of events. We have at our disposal a more powerful array of tools for mapping genes and QTLs by placing SNPs or STRs equidistantly throughout the genome and locating their positions through linkage analyses and association studies. We are now experiencing the repercussions of a molecular revolution that has forever altered the field of anthropological genetics.

References


REFERENCES


REFERENCES


