

DNA Markers and Human Evolution

Rebecca L Cann, *University of Hawaii-Manoa, Honolulu, Hawaii, USA*

Genetic markers reveal nucleotide sequence diversity at different locations and frequencies within the human genome. Since not all loci respond to the same evolutionary forces, one goal of DNA marker use is to understand how information from disparate genomic fragments can generate a realistic understanding of human population changes in space and time.

Introduction

New mutations, when they arise in a population, are spread by natural selection, migration and genetic drift. New combinations of mutations along chromosomes are also assembled by genetic recombination, as sister chromatids synapse during meiosis. These forces have consequences for understanding the demography of human populations, and they alter the patterns of genetic variation in different groups.

Genetic markers can be thought of as simple characters of the genome that have multiple states at each character. Diploid humans can be homozygous or heterozygous, possessing one or two different states (alleles) per character, usually described as the locus. In sequence-level analysis, each nucleotide position can become a character, with the four different potential states (the two different purine or pyrimidine bases potentially present) reflecting for that site four possible alleles segregating in the whole population. When summed over the entire locus, the total number of potential alleles present in a population per locus is not fixed, but is a function of time and natural selection.

Populations with many alleles at a marker locus are said to be highly polymorphic, reflecting a threshold level of greater than 5% for any specific allele. Alleles found to be present less frequently are still useful; when scored as rare alleles, they allow researchers to track the movement of new mutations into populations. Markers are either analyzed separately as frequency data for a single locus, or combined with information for other loci nearby and presented as specific haplotypes in genetic linkage from an individual donor.

Marker Development

Genetic markers showed differences between human populations first by indirect methods, before researchers had the ability to easily scan the genome for deoxyribonucleic acid (DNA) states. One early example was the work of Mourant *et al.* (1976), who

Advanced article

Article contents

- Introduction
- Marker Development
- Marker Applications to Evolutionary Questions
- Highly Informative Loci
- Why Use Several Loci Together?

doi: 10.1038/npg.els.0005078

catalogued polymorphisms based on differences between blood groups and other serological characteristics. With the wide-scale adoption of electrophoretic techniques to separate molecules by size and net charge, proteins found in blood, liver and muscle made up the second information wave. These classical genetic markers often served as the starting point for applications with new DNA-based markers, in order to refine and reexamine ideas about the relationships between different populations. A broad survey by Cavalli-Sforza *et al.* (1994) took this approach, mapping differences in allele frequencies across continents. They sought patterns that would reflect ancient migrations, periods of population mixing and long periods of time in geographic isolation resulting from social or climatic disturbances that limited human dispersal capabilities in technologically simple societies.

When polymerase chain reaction (PCR) became a popular technique, the same population isolates were retested for what have now become the classical DNA-based markers (Mountain *et al.*, 1992). These markers include those genetic loci within the mitochondrial genome (mtDNA), markers on the Y chromosome and mutations within the insulin receptor and hemoglobin locus. While largely confirming many early studies based on classical markers, new DNA markers allowed researchers to compare the evolutionary forces that work differentially in men and women to alter the patterns of genetic variation found in humans.

Marker Applications to Evolutionary Questions

Contemporary human populations have been shaped by multiple evolutionary forces (Lahr and Foley, 1998). As a consequence, studies of specific regions of the human genome must disentangle the confounding effects of these forces, distinguishing

between patterns caused by natural selection and patterns caused by changes in population size. At a locus where little variation is found to occur, how does one distinguish between the force of genetic drift, which reduces variation to just a few or one allele per locus over time, and the rare mutational event that occurred only recently, so that the probability of including in a survey that unique individual or its few descendants, who also carry the new mutation, is extremely improbable no matter how the population was sampled?

The most effective analysis of DNA markers incorporates the disciplines of linguistics as well as biological anthropology, in order to interpret the significance of particular changes in allele frequency over time. All markers are thought to ultimately reflect differences in DNA sequences, and population studies that use genetic markers, although not originally DNA-based, have long provided insights about connections among ethnic, religious and geographically distinct isolates.

Mutations leave their trail of information in the genome. They can be followed using multiple genetic systems from the same DNA sample. Processes that affect individuals through mutation and genetic recombination ultimately lead to the accumulation of differences in populations. These differences can contribute to structures within and between populations. Kinship systems, infectious disease epidemics, social customs and other rules about mate choice further enhance these differences. Some societies have enforced particular rules about mate choice for many years, and the allele frequencies for certain markers show unusual patterns that reflect these biases. Warfare that results in males being killed but females incorporated as slaves will lead to different patterns of genetic markers for the maternal and paternal genetic contributions. Even with these expectations, it is necessary to search through large amounts of sequence to find informative markers, as two randomly chosen humans are more likely to be identical to each other at a given locus than to have differences in their DNA sequences.

Composite genotypes of multiple loci can now also be assembled on microarrays and used to address issues about the level of inbreeding or fitness of individuals with specific attributes. Since these genotypes will be destroyed in a single round of sexual recombination, however, evolutionary biologists have not focused on their usefulness for reconstructing genealogies. Instead, these genotypes are shown to facilitate the use of genetic markers in forensics where the DNA available for testing may be so rare or so fragmented that PCR is impossible. They are also useful for molecular diagnostics of infectious disease and markers associated with pathogen resistance.

Highly Informative Loci

Populations under study might have diverged only in the last few thousand years from a common ancestral gene pool, so there may have been too few rounds of DNA replication, followed by recombination and repair, to generate many differences. Considering the amount of absolute time that may have elapsed since two groups began diverging is essential for designing genetic screens of populations (Sunnucks, 2000). Too little information will result in no signal, if an inappropriate marker is chosen. In order to be judged highly informative, a region should be characterized by a rapid mutation/recombination rate balanced against a normal or low DNA repair rate. These two conditions then promote a situation where multiple alleles at the same locus can segregate within the population. In the extreme, it is also possible to have too much information, because if populations appear so different, it may be impossible to link them through intermediate states or distinguish between back mutations and multiple substitutions affecting the same marker.

Sensitivity to the question of temporal scale is especially important in human evolutionary studies, because our shared common and recent history as a small population implies that few loci will contain enough variation to generate good DNA markers. Since regions of the genome can have hot spots (for instance, a location near several transposable elements) as well as cool spots (in centromeric heterochromatin), there is great emphasis placed on finding genomic regions with sufficient differences to be informative in population studies. A common strategy to devise new informative markers is to design PCR primers that anneal to conserved portions of exons and span an intron of 300–500 nt in size, followed by cloning and analysis of the variety of sequences seen within that region. Introns that vary in size can be detected on simple agarose gels, and introns that vary in sequence may contain specific restriction endonuclease recognition sites.

Some nuclear gene families, such as the HLA locus or the major histocompatibility complex (MHC) superfamily on human chromosome 6, are characterized as both recombination and mutational hot spots and also contain loci that are highly informative (Ayala, 1996). MHC class 2 genes are typically screened for mutations in exon 2, where natural selection may have promoted nucleotide change to generate a variety of different antigenic binding sites. However, the detection system used in a genomic screen with the MHC needs to be able to resolve a large number of heterozygotes, which are expected because there are so many different alleles potentially segregating that include duplicate functional copies

and pseudogenes. Each individual's diploid genome contains a maximum of two states per character, but is the exact same marker present in two different individuals? Specific amplification of the desired product may be impossible, and markers need to be fully characterized at the level of DNA sequencing to answer that question and to be useful in constructing gene genealogies.

Other systems, such as the maternally inherited mtDNA loci, are mutationally labile and nonrecombining. Human mtDNA has a major noncoding region of approximately 1100 nt and contains 37 genes, but none have introns. The genome is circular, and all 37 genes are inherited as a linked haplotype, with a range of substitution rates across functional regions. The advantage, compared with a diploid nuclear locus, is that cloning is not necessary to distinguish heterozygotes. Using either sequences from the entire human mtDNA genome or hypervariable portions within the major noncoding region (Ingman *et al.*, 2000), along with the aid of appropriate computer programs, researchers obtained a reconstructed family tree with the most recent common ancestor (MRCA) of our common mother's mtDNA sequence. This phylogenetic tree placed her within a sub-Saharan population from East Africa.

An analogous advantage occurs when utilizing Y-chromosome DNA markers. In the case of the Y, however, repetitive sequences limited early gene mapping and sequencing studies to a handful of markers in regions near testis determining genes. Because these markers were mostly in conserved regions, they yielded little information about population diversity. One informative marker was associated with the presence or absence of an *Alu* repeat, a short interspersed repetitive element (SINE), mapped to the Y chromosome of certain men (Hammer *et al.*, 2001). Gradually, conserved primers for amplifying adjacent and specific variable Y regions were published, and it became possible to assay these regions via PCR to look for length and sequence variation. Nonrecombining regions of the human Y chromosome now contribute to population studies with highly informative markers, and allow geneticists to characterize specific paternal versus maternal inputs into a population.

Single nucleotide polymorphisms (SNPs) are another common type of genetic variation often used to uncover hidden patterns of population differences. Microchip technology has allowed sets of SNPs to be scored for haplotype analysis of mutations in specific cell types, as well as from donors of different ethnic groups. DNA sequencing of homologous blocks of loci from multiple donors first revealed these markers in gene regions that corresponded to the exons, introns and flanking regions around specific genes. Denaturing high-performance liquid chromatography (DHPLC) is

now used as a technical innovation to avoid direct sequencing and reveal additional SNPs, usually after some initial round of DNA sequencing had drawn attention to genomic regions characterized by high variability. This approach allowed Underhill *et al.* (2000) to enlarge the number of informative Y markers. A phylogenetic tree based on 167 DNA markers for the Y chromosome of 1062 unrelated men now places their last common paternal ancestor in Africa.

Why Use Several Loci Together?

Multilocus genotyping utilizing SNPs, as well as randomly amplified polymorphic DNA (RAPDs), variable number of tandem repeats (VNTRs), amplified fragment length polymorphic DNA (AFLP) or single locus techniques (microsatellites or single-copy nuclear (scn) DNA regions), have also been used to conduct population-level screens of the human gene pool (Linares, 1999). These markers are usually scored by nucleotide state, or bands on electrophoretic gels. They may be expressed as simple dominants, present or absent in the individual screened, or given names corresponding to the specific size of a repeat at that locus.

Multilocus methods are attractive because broad regions of the genome can be sampled quickly. However, data are not always comparable between studies, and early surveys of the human gene pool with these markers had a noted ascertainment bias, since regions to be screened were initially identified based on variation in just a subset of a particular group of DNA donors. There are also biological limitations to such techniques, because the variation scored may not be heritable and might even be due to an infection of the genome by another organism, such as a virus, that was subsequently integrated. Duplicate, nested PCR reactions can address this problem. In spite of these limitations, PCR-based multilocus typing represents an economical alternative to sequence-based surveys of very large populations.

Short tandem repeats (STRs) used in multilocus typing are a special class of markers that vary in length among individuals. They are especially helpful when amplified from defined genomic regions based on sequence-tagged sites (STSs) used in positional cloning, because both the chromosome from which they are derived and their precise location relative to other markers in the human genomes physical map are known. If frequencies of STR markers in a population depart from Hardy–Weinberg equilibrium conditions, it is possible to examine their linkage state with functional loci that are close by. This is important, because many models in genetics dealing

with correlated variables assume that in an ideal population (infinitely large, with random mating), linkage of a new allele with its neighboring markers decays to zero in approximately 10 generations. In practice, however, we have discovered that linkage may persist for thousands of generations, owing to the force of natural selection.

There are three principal reasons why multilocus approaches are valuable in DNA marker studies. First, they allow for detection of characters that are nonconcordant. When populations of individuals show two different patterns of variation involving the same regions of the genome, the researcher is likely to discover something interesting in the biology of that organism. Sexual selection may distort the frequency with which particular loci are associated with mating success, so that markers on the Y chromosome might be expected to reflect different patterns of genetic variation compared with those on autosomes. Second, when rates of mutation and fixation (scored together as a substitution rate) vary over several orders of magnitude as a result of stochastic causes, multilocus sampling will help minimize bias caused by using markers with unusual and unknown mutation/recombination characteristics. Finally, combining data from several loci also gets around the criticism that there is danger in trying to describe the history of a population based on the history of a single gene, or a single linked set of genes.

See also

Gene Trees and Species Trees
Great Apes: Phylogenetics
Human Populations: Evolution
Phylogenetics

References

- Ayala FJ (1996) HLA sequence polymorphism and the origin of humans. *Science* **270**: 1930–1936.
- Cavalli-Sforza LL, Menozzi P and Piazza A (1994) *The History and Geography of Human Genes*. Princeton, NJ: Princeton University Press.
- Hammer MF, *et al.* (2001) Hierarchical patterns of global human Y-chromosome diversity. *Molecular Biology and Evolution* **18**: 1189–1203.
- Ingman M, Kaessmann H, Paabo S and Gyllensten U (2000) Mitochondrial genome variation and the origin of modern humans. *Nature* **408**: 708–713.
- Lahr MM and Foley RA (1998) Towards a theory of modern human origins: geography, demography, and diversity in recent human evolution. *Yearbook of Physical Anthropology* **41**: 137–176.
- Linares AR (1999) Microsatellites and the reconstruction of the history of human populations. In: Goldstein DB and Schlotterer C (eds.) *Microsatellites, Evolution and Applications*, pp. 183–197. London, UK: Oxford University Press.
- Mountain JL, Lin AA, Bowcock AM and Cavalli-Sforza LL (1992) Evolution of modern humans: evidence from nuclear DNA polymorphism. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **337**: 159–165.
- Mourant AE, Kopec AC and Domaniewska-Sobczak K (1976) *The Distribution of the Human Blood Groups and Other Polymorphisms*, 2nd edn. Oxford, UK: Oxford University Press.
- Sunnucks P (2000) Efficient genetic markers for population biology. *Trends in Ecology and Evolution* **15**: 199–203.
- Underhill PA, Shen P, Lin A, *et al.* (2000) Y chromosome sequence variation and the history of human populations. *Nature Genetics* **26**: 358–361.

Further Reading

- Cann RL (2001) Genetic clues to dispersal in human populations: retracing the past from the present. *Science* **291**: 1742–1748.
- Olson S (2002) *Mapping Human History*. Boston, USA: Houghton Mifflin Company.
- Przeworski M, Hudson RR and Di Renzo A (2000) Adjusting the focus on human variation. *Trends in Genetics* **16**: 296–302.
- Tishkoff SA and Williams SM (2002) Genetic analysis of African populations: human evolution and complex disease. *Nature Reviews Genetics* **3**: 611–621.