DECIPHERING POPULATION AFFINITY AND SUBSTRUCTURE USING Y CHROMOSOME SNP HAPLOTYPES

Peter A. Underhill
Department of Genetics, Stanford University, Stanford, CA

ABSTRACT

Binary DNA sequence variants like single nucleotide substitutions (SNPs) and small insertions or deletions (DIPs) associated with the non-recombining portion of the haploid Y chromosome provide a unique metric into population characterization and substructure. Progress in the identification of hundreds of PCR compatible binary Y polymorphisms in diverse populations has created new opportunities for male specific DNA analysis including providing clues to biogeographical ancestry.

INTRODUCTION

The paternally inherited haploid Y chromosome comprises the single largest non-recombining component of the human genome. Most of the molecule escapes the scrambling effects of recombination and represents the largest haplotypic region of the human genome. The sequential accumulation of single nucleotide substitutions (SNPs) and small insertions or deletions (DIPs) across the generations, can be determined. Generally for these types of data usually only two alleles are observed. Using the principles of maximum parsimony an evolutionary stable genealogy is created in which a network of branches is drawn that minimizes the number of mutational events that relates the lineages. Every Y chromosome can be assigned to a known branch in the tree that defines a haplotype. A haplotype is an array of specific alleles on a single chromosome analogous to a binary “0’s (ancestral allele) and 1’s” (derived allele) bar code. What is central is the assumption that the derived allele arose once in human history, and all men that display a particular mutant allele descend from a common ancestor on which the mutation first appeared. Informally, the last known mutation to occur on a particular chromosome can be used to define a particular lineage. In a similar manner, mitochondrial DNA (mtDNA), provides the analogous female record, although it’s inherently higher mutation rate causes greater recurrence and reversion and thus displays more phylogenetic “noise”. The lower effective population size for the Y chromosome in the overall gene pool translates into increased levels of population subdivision respective to other DNA sequences. The rarity of back and recurrent mutations further contributes to the property of displaying the strongest geographic correlation and greatest diversity amongst geographically distinct populations. Accordingly, the ability of molecular genetics to exquisitely define various clades of Y-chromosomes with distinctive phylogeographic character provides considerable transparency concerning a sample’s paternal biogeographical ancestry.

PROGRESS

The application of denaturing high performance liquid chromatography (1) to efficiently search for DNA sequence variation has transformed the landscape of this molecule from an impoverished locus (2) to one of considerable polymorphic richness. Since only a small fraction of the molecule has been surveyed thus far and hundreds of variants already detected (2) thousands of as yet uncharacterized polymorphisms potentially exist. This makes it feasible to undertake targeted searches on existing categories of known haplotypes to improve phylogenetic resolution. Since contiguous regions are typically surveyed by DHPLC during the marker discovery phase of the research, the polymorphisms discovered are often clustered sufficiently close such that several sites can be amplified simultaneously within a span of a few kilobases (Fig 1). Since we have focused on surveying single copy regions rather than repetitive sequences in the development of our inventory of hundreds of Y binary polymorphisms, observation of only one or the other allele at each marker is normally expected. When using such markers there will be no heterozyote signal to detect other than in mixtures of unrelated males.

GENEALOGY AND NOMENCLATURE

The determination of the key architectural elements of the Y chromosome genealogy (3) has revealed the basic framework on which further molecular resolution can be built. Recently the Y Chromosome Consortium (YCC) initiated a flexible nomenclature that accommodates future progress and overrides and unifies previous systems (4). The YCC nomenclature can be accessed at http://ycc.biosci.arizona.edu/nomenclature_system/fig1.html. This practical development represents another step in the maturation of the understanding the pattern of DNA sequence variation on this molecule.

PHYLOGEOGRAPHY

The absence of recombination and haploid nature of the Y chromosome permits the reconstruction of an unequivocal haplotype phylogeny based on the geographic distribution of the Y chromosome binary chromosomes, an approach known as phylogeography. The underlying assumption of phylogeography is that there is a correspondence between the overall distribution of haplotypes and haplogroups and past human movements. The strong geographical signal seen in the Y chromosome data is consistent with this assumption (5). While numerous factors converge to reduce the effective size of human Y-chromosomes in the gene pool relative to the autosomes, the consequence is to create the general situation whereby greater Y chromosome diversity exists between populations than among populations. The correlation of Y chromosome haplotypes with geography is illustrated in Fig 2. Such patterns provide potentially important clues to paternal biogeographical ancestry.

The non-random distribution of Y chromosome haplotypes observed in Europe has been proposed to be mainly a function of re-colonization episodes by descendents of earlier Paleolithic foragers from various refugia following the Last Glacial Maximum (6). One of these haplotypes is defined by M170 that appears to have arisen in situ within Europe and is thus an excellent indicator of European paternal heritage and gene flow. The frequency distribution of M170 related lineages is shown in Fig 3.

USES OF Y CHROMOSOME BINARY MARKERS

Determining the haplotypes of the non-pseudoautosomal region of the Y chromosome can be useful in molecular forensics (7, 8). While molecular forensics usually focuses on individual genetic profiles, informative population based signatures now provide a straightforward simple tool for exclusion. Panels of binary markers for haplotype determination and amenable to various genotyping technologies can be tailored to either encompass a broad geographic spectrum or a more localized geographical ancestry. Y chromosome data can be used to differentiate mixtures of male and female contributors. They can also be used to distinguish sibs from half sibs as well as identify multiple unrelated male perpetrators by reconstruction haplotypes from apparent heterozygote allele calls. These markers and the haplotypes they define will be valuable reagents for genealogical studies and illuminate paternal biogeographical ancestry.

CAVEATS

The Y chromosome markers are limited only to males. Despite a very low nuclear mutation rate, a low percentage (ca. 1%) of recurrent and reversion mutation has been observed. Fortunately the amount of mutations now available to define haplotypes by positive character state within the genealogy makes detection of recurrent and reversion mutation events easily and unequivocally recognized. Since the Y chromosome is a single locus it is possible that Y chromosome ancestry may not always reflect the ancestry of the rest of the genome, such haplotypes.

SUMMARY

The forensic community is justifiably conservative and risk adverse. Nonetheless, sufficient molecular and population genetic knowledge and intellectual infrastructure now exists for the introduction and use of Y

chromosome binary markers within molecular forensic science. While practical issues remain regarding marker selection and the choice(s) of simple, robust genotyping approaches, it is a propitious time to begin applying the unique and informative properties of Y chromosome binary polymorphisms to the field of human molecular genetic identity.

REFERENCES

1. Underhill, P. A., Jin, L., Lin, A. A., Mehdi, S. Q., Jenkins, T., Vollrath, D., Davis, R. W., Cavalli-Sforza, L. L., and Oefner, P. J. Detection of numerous Y chromosome biallelic polymorphisms by denaturing high performance liquid chromatography (DHPLC). Genome Res. 1997, 7: 996-1005.

2. Jobling M. A., Tyler-Smith C. Fathers and sons: the Y chromosome and human evolution. Trends in Genetics 1995, 11:449-456.

3. Underhill, P. A., Shen, P., Lin, A. A., Jin, L., Passarino, G., Yang, W. H., Kauffman, E., Bonné-Tamir, B., Bertranpetit, J., Francalacci, P., Ibrahim, M., Jenkins, T., Kidd, J. R., Mehdi, S. Q., Seielstad, M. T., Wells, R. S., Piazza, A., Davis, R. W., Feldman, M. W., Cavalli-Sforza, L. L., and Oefner, P. J. Y chromosome sequence variation and the history of human populations. Nat. Genet. 2000 **26**: 358-361.

4. The Y Chromosome Consortium. A Nomenclature System for the Tree of Human Y-Chromosomal Binary Haplogroups. Genome Res. 2002, 12: 339-348.

5. Underhill, P. A., Passarino, G., Lin, A. A., Shen, P., Foley, R. A., Mirazón Lahr, M., Oefner, P. J. and Cavalli-Sforza, L. L. The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. Ann. Hum Genet 2001. 65: 43-62.

6. Semino, O., Passarino, G., Oefner, P. J., Lin, A. A., Arbuzova, S., Beckman, L. E., De Benedictis, G., Francalacci, P., Kouvatsi, A., Limborska, S., Marcikiæ, M., Mika, A., Mika, B., Primorac, D., Santachiara-Benerecetti, A. Cavalli-Sforza, L. L. and Underhill, P. A. The genetic legacy of Palaeolithic *Homo sapiens sapiens* in extant Europeans: a Y-chromosome perspective. Science 2000**,** 290: 1155-1159.

7. National Institute of Justice. A report from the National Commission on the future of DNA evidence. The Future of Forensic DNA Testing: Predictions of the Research and Development Working Group. November 2000, NCJ 183697. 91 p.

8. Jobling, M. A. Y-chromosomal SNP haplotype diversity in forensic analysis. Forensic Sci. International 2001. 118: 158-162.

FIGURE LEGENDS

Fig. 1. Diagram of a single copy region of the human Y chromosome showing the physical distribution of various binary markers. Clusters of markers successfully amplified are indicated

Fig. 2. Worldwide distribution of 10 clusters of Y chromosome haplotypes in 22 geographic regions. Reproduced from reference 5.

Fig. 3. The frequency distribution of M170 related Y chromosome lineages in Europe. Data taken from reference 6.