DHPLC Applications: Finding DNA variation on the Y chromosome

Qasim Ayub1, Aisha Mohyuddin1, Shagufta Khaliq1, Abdul Hameed1, Chris Tyler-Smith2 and S. Qasim Mehdi1

1Biomedical and Genetic Engineering Division, Dr. A. Q. Khan Research Laboratories, P.O. Box 2891, Islamabad 44000, Pakistan.
2CRC Chromosome Molecular Biology Group, Department of Biochemistry, University of Oxford, Oxford, UK.

Abstract
Denaturing High-Performance Liquid Chromatography (DHPLC) is a recently developed technique for the detection of single nucleotide polymorphisms (SNPs) and mutations. It involves the comparison between two or more DNAs as a mixture of denatured and reannealed PCR products. The methodology is based on the principle of reversed phase liquid chromatography and uses a unique DNA separation matrix. The exquisite sensitivity of the technique is determined by adjusting the oven temperature. Elution of DNA fragments is dependant on the chain length and sequence and can be predicted by computation. Under partially denaturing conditions heteroduplexes formed upon mixing, denaturing, and reannealing of two, or more, chromosomes that differ in sequence, are retained less than their corresponding homoduplexes and sequence variation is recognized by the appearance of two, or more, peaks in the chromatographs. Numerous SNPs have been identified on the non-recombiant portion of the Y chromosome by using this technique. To investigate the DNA variation within Pakistan 15 Y-SNPs, an Alu insertion, a LINE1 insertion and the 12f2 deletion, mapping on the non-recombining portion of the human Y chromosome, were typed in 834 Pakistani males. The combination of these biallelic markers identified 11 stable Y chromosomal lineages or Y ‘haplogroups’ in the Pakistani population. Haplogroup frequencies were generally similar to those in neighboring geographical areas and indicate that there was a common pool of Y lineages within Pakistan that are predominantly from West and Central Asia.

Keywords: DHPLC, Y-polymorphisms, Y SNPs, Population genetics, Pakistan.

INTRODUCTION
Denaturing High-Performance Liquid Chromatography (DHPLC) is a recently developed technique for the detection of single nucleotide polymorphisms (SNPs) and mutations. It involves the comparison between two or more DNAs as a mixture of denatured and reannealed PCR products. The PCR products do not require purification and single nucleotide substitutions, deletions and insertions have been detected successfully by on line monitoring within 2-6 minutes (For review see Xiao and Oefner 2001). The instrumentation is based on the principle of ion-paired reversed phase high-performance liquid chromatography and uses a unique DNA separation matrix. The proprietary matrix (DNASep column) consists of an alkylated non-porous poly-styrene divinyl benzene C-18 particles, 2-3 μm in diameter. The double stranded DNA molecules are retained on the column by binding with amphiphilic ions (0.1 M tri-ethylammonium acetate (TEAA) buffer, pH 7.0. The PCR products are eluted in 0.1 M TEAA buffer (pH 7.0) with a linear acetonitrile gradient at a flow rate of 0.9 ml/min. The exquisite sensitivity of the
technique is determined by adjusting the oven temperature. The optimum temperature and elution profiles are dependant upon the chain length and sequence and can be predicted by computation (http://insertion.stanford.edu/melt1.html). Under partially denaturing conditions heteroduplexes formed upon mixing, denaturing, and reannealing of two, or more, DNAs that differ in sequence, are retained less than their corresponding homoduplexes and sequence variation is recognized by the appearance of two, or more, peaks in the elution profiles (Figure 1). This methodology has been successfully applied for high throughput screening and identification of numerous Y chromosomal SNPs (Underhill et al. 1997, 2001). Briefly sequences of the non-recombinant portion of the human Y chromosome available in public databases (such as http://www.ncbi.nlm.nih.gov/) were screened In silico for unique sequences. Human repeat DNA sequences were identified by masking repeat sequences using the RepeatMasker2 software (http://ftp genome.washington.edu/cgi-bin/RM2).

Primers were designed (http://www-genome.wi.mit.edu/genome_software/-other/primer3.html) to amplify 150-700 base pairs of unique sequences by PCR. Heteroduplex analyses were carried out by mixing, denaturing and reannealing of unpurified PCR product from a single reference male individual with an amplicon from another individual (Figure 1). Y-SNPs were identified by the appearance of two or more peaks in the elution profiles and confirmed by DNA sequencing.

The Y chromosome carries the largest non-recombining segment in the human genome and this allows investigation of the male genetic history. To investigate the DNA variation within Pakistan 18 biallelic markers mapping on the non-recombining portion of the human Y chromosome were examined. These included 5 Y-SNPs that had been identified previously by DHPLC (Underhill et al. 1997).

MATERIALS AND METHODS

Samples: The Y chromosomes of 834 unrelated Pakistani males belonging to fifteen ethnic groups of Pakistan were analyzed. The details about the various ethnic groups are given elsewhere (Mehdi et al. 1999). Informed consent was obtained from all participants in this study. DNA was extracted from EBV-transformed lymphoblastoid cell lines that were established for each individual.

Y-Chromosome Binary Polymorphisms: Eighteen biallelic markers, including 15 Y-SNPs, an Alu insertion (Hammer 1994; Hammer and Horai 1995), a LINE1 insertion (Santos et al. 2000) and the 12f2 deletion (Casanova et al. 1985), were examined. The SNPs studied were 92R7 C to T (Mathias et al. 1994); SRY-2627 C to T (Bianchi et al. 1997); SRY-1532 A to G to A (Kwok et al. 1996; Santos et al. 1999b; Whitfield et al. 1995); sY81 A to G (Seielstad et al. 1994); SRY-8299 G to A (Santos et al. 1999a); Apt G to A (Pandya et al. 1998); SRY +465 C to T (Shinka et al. 1999); LLY22g C to A, Tat T to C transition (Zerjal et al. 1997) and the RPS4Y C to T mutation (Bergen et al. 1999). Markers that were identified by DHPLC included M9, M11, M17, M20 and M48 (Underhill et al. 1997). The set of Y binary marker alleles carried by a single individual was referred to as the ‘haplogroup’ (Tyler-Smith 1999).

Data Analysis: Principal components analysis was carried out on haplogroup frequencies using the ViSta (Visual Statistics) system software version 6.0 (Young and Bann 1996). For graphic representation the first and second principal components were plotted using the Microsoft Office Suite Excel Package on Windows 2000.
RESULTS

The combination of the biallelic markers identified 11 stable Y chromosomal lineages or Y ‘haplogroups’ in the Pakistani population (Figure 2). Of these five haplogroups (1, 2, 3, 9 and 28) accounted for 92.1% of the sample. Haplogroups 1, 3 and 9 were present in all Pakistani populations examined except for the Mohanna, Hazara and Punjabi respectively. Haplogroup 28 was found in most Pakistani ethnic groups except for the Hazara, Kashmiri and Punjabi. The south-western Pakistani populations (Baloch, Brahui, Makrani Baloch, Makrani Negroid, Parsi, and Sindhi) show higher frequencies of haplogroup 9 and the YAP+ haplogroups 21 and 8 compared to the northern populations (Balti, Burusho, Hazara, Kalash, Kashmiri, Pathan, Punjabi). However, no strong overall geographical clustering of haplogroup frequencies is apparent within the country.

The principal components analysis (Figure 3) shows that the Pakistani populations cluster with each other except for the Hazara. It is striking that the Pakistani populations speaking a language isolate

---

**Figure 2. Phylogeny of Y haplogroups in Pakistan.** An unrooted maximum parsimony tree showing Y haplogroups (numbers in circles) and mutations (on lines). Circle area is proportional to haplogroup frequency in Pakistan.

**Figure 3. Principal components analysis of Y haplogroup frequencies within Pakistan.** Indo-European speakers are indicated by triangles, Sino-Tibetan by a filled triangle, Dravidian by a circle and the language isolate Burusho by a diamond.
(the Burushos), a Dravidian language (the Brahui), or a Sino-Tibetan language (the Balti) resembled the Indo-European-speaking majority and do not stand out in these analyses.

DISCUSSION

To investigate the DNA variation within Pakistan 15 Y-SNPs, an Alu insertion (Hammer 1994), a LINE1 insertion (Santos et al. 2000) and the 12f2 deletion (Casanova et al. 1985) were examined in 834 male samples. Of these 5 SNPs had been identified by DHPLC. This technique has been successfully used to identify numerous such polymorphisms on the non-recombinant part of the human Y chromosome (Underhill et al. 1997; 2001).

The combination of these markers identified 11 stable Y chromosomal lineages or Y ‘haplogroups’ within Pakistan. The haplogroup frequencies within Pakistan were generally similar to those in neighboring geographical areas to the west and the Pakistani populations speaking a language isolate (the Burushos), a Dravidian language (the Brahui), or a Sino-Tibetan language (the Balti) resembled the Indo-European-speaking majority. The Hazaras are the most distinct among the Pakistani populations.

They are Mongoloid and claim descent from Genghis Khan’s army. Their Y chromosomal DNA variation supports this claim. The greater genetic similarity of Pakistani populations to those in the west than to eastern populations is illustrated by the fact that four of the five frequent haplogroups in Pakistan (1, 2, 3 and 9) are also frequent in Western Asia and Europe but not in the Chinese or Japanese. These patterns can be accounted for by a common pool of Y lineages within Pakistan that arrived predominantly from West Asia and the Levant (Qamar et al. 1999, Qamar et al. 2002; Quintana-Murci et al. 2001). This is supported by historical records that describe the arrival of the Indo-European language speaking pastoral nomads from Western and Central Asia into present day Pakistan (Wolpert 2000). The fifth common haplogroup in Pakistan is haplogroup 28. It is delineated by the M20 A to G transition and was originally identified by DHPLC in a Pakistani sample (Underhill et al. 1997). This haplogroup makes up to 13% of the Pakistani sample. It has rarely been found outside Pakistan and is probably associated with the spread of agriculture in the Indus Valley (Qamar et al. 2002). The use of DHPLC technology has greatly facilitated studies on human migration and evolution from the male perspective and shed light on the origins of modern human populations.

Acknowledgments

This work was supported by a Wellcome Trust Collaborative Research Initiative Grant to SQM. We express our gratitude to the original DNA donors who made this study possible.

References


