International Journal of Biology; Vol. 7, No. 1; 2015
ISSN 1916-9671 E-ISSN 1916-968X
Published by Canadian Center of Science and Education

Inferring Human Phylogenies Using Three CODIS STR Markers (CSF1PO, TPOX and TH01)

Nuzhat A. Akram[1] & Shakeel R. Farooqi[1]

[1] Department of Genetics, University of Karachi, University Road, Karachi 75270, Pakistan

Correspondence: Shakeel R. Farooqi, Department of Genetics, University of Karachi, University Road, Karachi 75270, Pakistan. Tel: 92-21-9926-1388, 923002589170. E-mail: farooqis@uok.edu.pk, smr.akram@gmail.com

Received: August 25, 2014 Accepted: September 8, 2014 Online Published: October 14, 2014

doi:10.5539/ijb.v7n1p1 URL: http://dx.doi.org/10.5539/ijb.v7n1p1

Abstract

Over the past several decades polymorphic genetic loci have been discussed for their utility in human phylogenetic inferences. Short Tandem Repeat (STR) loci have shown promising results for this purpose. Unfortunately, allele frequency data of polymorphic loci are largely confined to few populations. Therefore, the number of shared loci declines as the number of population increases. We hypothesize that even a smaller number of STR loci can be used efficiently for phylogenetic purposes if an appropriate theoretical and statistical strategy is employed. This strategy provides a feasible and cost effective method to choose appropriate STR loci for phylogenetic studies. For this purpose, an empirical study was conducted using allele frequency data of three STR loci CSF1PO, TPOX, and TH01 across 98 human populations from the literature (references are available at http://dnaa.bravehost.com/index.html and http://www.cstl.nist.gov/strbase/population/Omnipop). The choice of markers was based on locus polymorphism, high heterozygosity, low mutation rate, less artifacts and independence between the loci. Three methods were used to measure genetic distances between the populations; Cavalli Sforza's chord distance (D_C), Nei's genetic (D_A) and Nei's standard genetic distances (D_{ST}). Coefficient of variation (CV) was calculated across hundred (100) datasets obtained by re-sampling of the original dataset for each of the genetic distance methods. CV was in order of $D_{ST} > D_A > D_C$. Therefore, a consensus tree based on D_C was constructed using Neighbour Joining (NJ), Unweighted Pair Group Method with Arithmatic mean (UPGMA) and Maximum Likelihood (ML) methods. NJ and UPGMA methods got more statistical support that is higher bootstrap values than ML (NJ> UPGMA> ML). Validation study was performed using (A) Principal Component Analysis (B) Comparison with trees reported for other molecular markers (C) STR genotyping of five Pakistani subpopulations. Results strongly supported our hypothesis that the three STR markers CSF1PO, TPOX, and TH01 are successful in delineating ethnic, geographic and linguistic differentiation between the populations.

Keywords: human phylogenetic inferences, STR loci *CSF1PO TPOX TH01*, world population data, Pakistani subpopulations.

1. Introduction

Phylogenetic inferences are premised on the inheritance of ancestral characteristics and on the existence of an evolutionary history defined by changes in these characteristics (Li, Pearl, & Doss, 2000). Indeed, many human populations carry distinct genetic markers, and by tracing these markers through the generations their origin can be traced out (Adams, 2008). Since many decades allele frequency data have been used to reconstruct evolutionary histories of human populations (Ayub et al., 2003; Agrawal & Khan, 2005). A number of statistical problems related to the number of loci, sample size of the populations, degree of locus/loci polymorphism, distance methods, methods of reconstructing phylogenetic trees and the methods to ensure the reliability of the trees…etc have been addressed in the literature (Zhivotovsky & Feldman, 1995; Takezaki & Nei, 1996; Nei & Takezaki, 1996; Felsenstein, 2003; Holder & Lewis, 2003; Takezaki & Nei, 2008). However, polymorphic loci for which the allele frequency data are available are largely confined to European, North American and East Asian populations (Nei & Roychoudhury, 1993). For this reason the number of shared loci declines as the number of population increases. Therefore if one wants to use a large number of loci, the number of populations that can be used becomes very small. Moreover, the missing elements in locus × population matrix often introduce unreasonable branching patterns in phylogenetic trees. In the present study we hypothesize that the minimum number of markers can perform efficiently for phylogenetic inferences if the theoretical and statistical strategy applied is correct. For this

purpose three microsatellite loci, also called Short Tandem Repeats (STRs), were chosen from Combined DNA Index System (CODIS) of American Federal Bureau of Investigation (Budowle, Moretti, Niezgoda, & Brown, 1998; Budowle, Moretti, Baumstark, Defenbaugh, & Keys, 1999; Butler, 2006). STRs are regions of tandemly repeated DNA segments found throughout the human genome that show length polymorphism with a core repeated DNA sequence (Butler & Hill, 2012). These markers are used for human identification purposes in forensic caseworks.

In the present work, an empirical study was conducted using STR allele frequency data of 98 human populations (references available at http://dnaa.bravehost.com/index.html and http://www.cstl.nist.gov/strbase/population/Omnipop). For the validation of the phylogenetic inferences five Pakistani subpopulations were genotyped for the three STR loci and the allele frequency data was incorporated into world population data. Statistical analyses performed on the datasets of empirical and validation studies are explained in section 'Theory and Calculations'.

2. Materials and Methods

2.1 Choosing the STR Loci

Three STR loci (CSF1PO, TPOX, and TH01) were chosen from Combined DNA Index System (CODIS) of American Federal Bureau of Investigation (FBI). Different chromosomal locations of these loci minimize the chances of linkage disequilibrium between them (Table 1). Moreover the variance of the number of heterozygous loci was found to be within 95% confidence interval which means there is no association between the loci under study. Two of the loci namely TPOX and TH01 showed the lowest mutation rate among all the CODIS STR loci, hence making them suitable for phylogenetic purposes (Keim et al., 2004). Frequency of biological "artifacts" associated with theses loci such as null alleles, stutter products, non-template nucleotide addition is low.

Table 1. Chromosomal locations and other information of the three STR loci (CSF1PO, TPOX and TH01)

S. No.	Locus Name	Locus Definition	Chromosomal Location	Repeat Sequence	Mutation Rate (%)
1	CSF1PO	Human c-fms proto-oncogene for CSF-1 receptor gene	5q33.3-34	AGAT	0.16
2	TPOX	Human thyroid peroxidase gene	2p25.1-pter	AATG	0.01
3	TH01	Human tyrosine hydroxylase gene	11p15.5	AATG	0.01

2.2 World Population Data

Allele frequency data of 62 world populations from the literature (http://dnaa.bravehost.com/index.html) and 36 world populations from the Omnipop excel file (http://www.cstl.nist.gov/strbase/population/Omnipop) were available to reconstruct the phylogenetic trees. These populations encompass major geographical areas of the world and show different ethnic and linguistic affiliations (Table 2 and Table 3). All three loci were reported to be in Hardy Weinberg equilibrium across all the populations under study.

Table 2. Ethnic, Linguistic and Geographic classification of the populations (http://dnaa.bravehost.com/index.html)

S. No.	Ref. No.[a]	Population/Subpopulation	Abbreviation used	Ethnic group	Language
1	37	African Jordanian	AfricanJ	Negroid	Niger-Congo
2	2	African Mozambique	AfricanM	Negroid	Niger-Congo
3	25	Andhra Pradesh, Golla caste1	APGolla1	Australoid	Indo-European
4	25	Andhra Pradesh, Golla caste2	APGolla2	Australoid	Indo- European
5	25	Andhra Pradesh, Golla caste3	APGolla3	Australoid	Indo- European
6	25	Andhra Pradesh, Golla caste4	APGolla4	Australoid	Indo-European
7	25	Andhra Pradesh, Golla caste5	APGolla5	Australoid	Indo- European
8	25	Andhra Pradesh, Golla caste6	APGolla6	Australoid	Indo-European
9	25	Andhra Pradesh, Golla caste7	APGolla7	Australoid	Indo-European
10	8	Bangladeshis	BNDeshis	Caucasoid	Indo-European
11	31	Bengal Tribe1	BengalT1	Mongoloid	Indo-European
12	31	Bengal Tribe2	BengalT2	Mongoloid	Indo- European

13	31	Bengal Tribe3	BengalT3	Mongoloid	Indo- European
14	31	Bengal Tribe4	BengalT4	Mongoloid	Indo- European
15	17	Bhutan	Bhutan	Mongoloid	Sino-Tibetan
16	35	Bohemian	Bohemian	Caucasoid	Indo-European
17	7	Bolivian	Bolivian	Caucasoid	Indo-European
18	11	Brazilian	Brazilian	Caucasoid	Indo-European
19	36	Brazilian2	Brazilian2	Caucasoid	Indo-European
20	27	Central India,Agharia	C.IndiaA	Caucasoid	Indo-European
21	27	Central India,Dheria Gond	C.IndiaD	Australoid	Indo-European
22	27	Central India,Satmani	C.IndiaS	Caucasoid	Indo-European
23	27	Central India,Teli	C.IndiaT	Caucasoid	Indo-European
24	15	Chinese	Chinese	Mongoloid	Austronesian
25	19	Chinese Hong Kong	ChnHkong	Mongoloid	Austronesian
26	38	Chinese Korean	ChnKorean	Mongoloid	Austronesian
27	1	Dubai, Bangladeshis	DubaiBang	Caucasoid	Indo- European
28	1	Dubai, Iranians	DubaiIran	Caucasoid	Afro-Asiatic
29	1	Dubai, Omanis	DubaiOman	Caucasoid	Afro-Asiatic
30	1	Dubai, Saudi Arabians	DubaiSArab	Caucasoid	Afro-Asiatic
31	1	Dubai, Yemenites	DubaiYemen	Caucasoid	Afro-Asiatic
32	9	Eastern India, Garo	E.IndiaG	Mongoloid	Indo-European
33	9	Eastern India,Brahmin	E.IndiaBr	Caucasoid	Indo-European
34	9	Eastern India,Kayastha	E.IndiaKa	Caucasoid	Indo-European
35	26	Ecuadorian	Ecuadorian	Caucasoid	Indo- European
36	6	Greek Cyprus	GrkCyprus	Caucasoid	Afro-Asiatic
37	34	Gurkha, Malaysia	GurkhaMLY	Caucasoid	Indo-European
38	3	India, Bihar Baniya	BiharBaniy	Caucasoid	Indo-European
39	3	India, Bihar Kurmi	BiharKurmi	Australoid	Indo-European
40	3	India, Bihar Yadav	BiharYadav	Caucasoid	Indo-European
41	4	India,Andhra Pradesh, Dravidian1	APDravidn1	Caucasoid	Dravidian
42	4	India,Andhra Pradesh, Dravidian2	APDravidn2	Caucasoid	Dravidian
43	4	India,Andhra Pradesh, Dravidian3	APDravidn3	Caucasoid	Dravidian
44	29	Iran	Iranian	Caucasoid	Afro-Asiatic
45	33	Japanese	Japanese	Mongoloid	Afro-Asiatic
46	24	Kashmiris	Kashmiris	Caucasoid	Indo-European
47	30	Kurd	Kurdish	Caucasoid	Afro-Asiatic
48	28	Malaysian Chinese	MLYchinese	Mongoloid	Afro-Asiatic
49	28	Malaysian Indians	MLYindians	Caucasoid	Indo-European
50	28	Malaysian Malays	MLYmalays	Mongoloid	Afro-Asiatic
51	10	Muslim Tamil Bohra	TamilBohra	Australoid	Dravidian
52	10	Muslim Tamil Sunni	TamilSunni	Australoid	Dravidian
53	16	Northern Greece	NGreece	Caucasoid	Indo-European
54	21	Singapore Indians	SPIndian	Caucasoid	Indo-European
55	12	South African Blacks	S.AfBlack	Negroid	Niger-Congo
56	12	South African Whites	S.AfWhite	Caucasoid	Indo-European
57	32	Thailand	Thailand	Mongoloid	Austronesian
58	20	Tibet Lassa	TibetLassa	Mongoloid	Austronesian
59	5	Zimbabwe(Black African)	Zimb. Bl. Af	Negroid	Niger-Congo
60	18	Nepal	Nepal	Mongoloid	Sino-Tibetan
61	23	South India Tamil	SIndTamil	Australoid	Dravidian
62	14	Bavarian Caucasians	Caucasians	Caucasoid	Indo-Europeans

[a] It refers to the number of reference provided at the website http://dnaa.bravehost.com/index.html

Table 3. Ethnic, Linguistic and Geographic classification of the populations (http://www.cstl.nist.gov/strbase/population/Omnipop)

S.No.	*Ref. No.*[a]	*Population/Subpopulation*	*Abbreviation used*	*Ethnic group*	*Language*
1	*1*	FBI African American	AfrAmer	Negroid	Niger-Congo
2	*2*	Bahama African American	AfrBahm	Negroid	Niger-Congo
3	*2*	Jamaica African American	AfrAmJm	Negroid	Niger-Congo
4	*2*	Trinidad African American	AfrAmTr	Negroid	Niger-Congo
5	*2*	California African American	AfrAmCa	Negroid	Niger-Congo
6	*2*	Alabama African American	AfrAmAL	Negroid	Niger-Congo
7	*2*	Florida African American	AfrAmFL	Negroid	Niger-Congo
8	*2*	Virginia African American	AfrAmVr	Negroid	Niger-Congo
9	*2*	New York African American	AfrAmNY	Negroid	Niger-Congo
10	*2*	Illinois African American	AfrAmIL	Negroid	Niger-Congo
11	*2*	Alabama Caucasians	CaucaAL	Caucasoid	Indo-European
12	*2*	Virginia Caucasians	CaucaVr	Caucasoid	Indo-European
13	*2*	Michigan Caucasians	CaucaMi	Caucasoid	Indo-European
14	*2*	Florida Hispanics	HispaFL	Caucasoid	Indo-European
15	*2*	Arizona Hispanics	HispaAr	Caucasoid	Indo-European
16	*2*	Chinese	Chinese	Mongoloid	Austronesian
17	*2*	Japanese	Japanes	Mongoloid	Afro-Asiatic
18	*2*	Korean	Koreans	Mongoloid	Austronesian
19	*2*	General Asians	GAsians	Caucasoid	Indo-European
20	*3*	Swiss Caucasians	CaucaSw	Caucasoid	Indo-European
21	*5*	Connecticut African American	AfrAmCo	Negroid	Niger-Congo
22	*5*	Connecticut Caucasians	CaucaCo	Caucasoid	Indo-European
23	*5*	Connecticut Hispanics	HispaCo	Caucasoid	Indo-European
24	*7*	Turkish	Turkish	Caucasoid	Indo-European
25	*9*	Southern Spain (Andalusia)	SSpaini	Caucasoid	Indo-European
26	*10*	Brazilian	Brazils	Caucasoid	Indo-European
27	*13*	Tamil (India)	TamilsI	Australoid	Dravidian
28	*2*	New York Caucasians	CaucaNY	Caucasoid	Indo-European
29	*6*	Basques	Basques	Caucasoid	Language Isolate
30	*56*	CFS Asian Canada	AsianCa	Caucasoid	Indo-European
31	*56*	CFS East Indian Canada	EIndian	Caucasoid	Indo-European
32	*58*	Desasthbrahmin (India)	DBrahmn	Caucasoid	Indo-European
33	*58*	Chitpavanbrahmin (India)	CBrahmn	Caucasoid	Indo-European
34	*59*	Tunisian	Tunisia	-	Afro-Asiatic
35	*61*	Oriya Brahmin (India)	OBrahmn	Caucasoid	Indo-European
36	*61*	Khandayat (India)	KIndian	Caucasoid	Indo-European

[a] It refers to the number of reference provided at the website (http://www.cstl.nist.gov/strbase/population/Omnipop).

2.3 Genotyping of the Three STR Loci (CSF1PO, TPOX and TH01) across Five Pakistani Subpopulations

One hundred and seventy five unrelated individuals (2n) were chosen from five Pakistani subpopulations residing in Karachi. Individuals within each subpopulation were selected through randomization. These subpopulations were Baloch (n = 64), Muhajir (Urdu speaking Indian immigrants) (n = 94), Pathan (n = 60), Punjabi (n = 74) and

Sindhi (n = 58). Each individual was genotyped for the three STR loci after taking informed consent. All the three loci were co amplified in a single PCR reaction using CTT (CSF1PO, TPOX and TH01) primers in 2400 thermal cycler. The protocols provided by the Promega Geneprint STR System Technical Manual (tm#004) were followed. Allele frequencies were estimated using maximum likelihood method (Li, 1976; Hedrick, 2011).

2.4 Theory and Calculations

2.4.1 STR Polymorphism

STR polymorphism was estimated by (i) *Heterozygosity (h).* Heterozygosity of each of the three loci was estimated by the Nei's unbiased formula

$$h = (2n/(2n\text{-}1))(1- \sum pi^2) \quad (1)$$

where *pi* is the frequency of *i*th allele in the sample, and *n* is the number of diploid individuals examined at the locus. The average Heterozygosity *(H)* of each of the three loci was estimated by

$$H = (\sum h)/s \quad (2)$$

where *s* is the total number of populations under study. (ii) Average number of alleles per locus (n_a). n_a was computed by the formula

$$n_a = (\sum N_a)/ s \quad (3)$$

where N_a is the number of alleles of a locus in a population and *s* is the total number of populations under study. (iii) Polymorphism Information Content (PIC) was calculated using the formula

$$PIC = 1- \sum pi^2 -2 \sum pi^2 pj^2, \quad (4)$$

where *pi* and *pj* stands for the frequencies of *i*th and *j*th alleles of a locus (Shete, Tiwari, & Elston, 2000; Kobilinsky, Liotti, & Oeser Sweat, 2005). (iv) Probability of Identity (PI) and Power of Discrimination (PD): PI is derived by the formula

$$PI = \sum (xi)^2 + \sum (xij)^2 \quad (5)$$

where *xi* stands for the frequency of homozygotes and is equal to pi^2. While *xij* stands for the frequency of heterozygotes and is equal to *2 pi pj*, where *pi* and *pj* stands for the frequencies of *i*th and *j*th alleles of a locus. PD is defined as,

$$PD = 1- \sum (xi)^2 + \sum (xij)^2 \quad \text{or} \quad 1- PI \quad (6)$$

Different measures/statistics of locus polymorphism across the world populations are shown in Table 4 and Table 5.

Table 4. Measures of locus polymorphism for the three STR loci (CSF1PO, TPOX and TH01) averaged over 62 world populations (http://dnaa.bravehost.com/index.html)

Measures of locus polymorphism	*Abbreviation used*	*CSF1PO*	*TPOX*	*TH01*
Average Heterozygosity	H	0.713	0.690	0.756
Average number of alleles	n_a	6.7	5.9	6.1
Polymorphism Information Content	PIC	0.690	0.619	0.708
Power of Discrimination	PD	0.837	0.798	0.849
Power of Identity	PI	0.163	0.202	0.151

Table 5. Measures of locus polymorphism for the three STR loci (CSF1PO, TPOX and TH01) averaged over 36 world populations (http://www.cstl.nist.gov/strbase/population/Omnipop)

Measures of locus polymorphism	*Abbreviation used*	*CSF1PO*	*TPOX*	*TH01*
Average Heterozygosity	H	0.739	0.689	0.763
Average number of alleles	n_a	10	10	10
Polymorphism Information Content	PIC	0.694	0.641	0.727
Power of Discrimination	PD	0.886	0.852	0.906
Power of Identity	PI	0.113	0.147	0.093

Heterozygosities of the three loci across each subpopulation of Pakistan are shown in Table 6.

Table 6. Observed heterozygosities of the three STR loci (CSF1PO, TPOX and TH01) across the five Pakistani subpopulations. These subpopulations were Baloch, Muhajir, Pathan, Punjabi and Sindhi

Pakistani Subpopulations	*CSF1PO*	*TPOX*	*TH01*
Baloch	0.531	0.625	0.875
Muhajir	0.532	0.851	0.787
Pathan	0.6	0.7	0.7
Punjbi	0.486	0.594	0.702
Sindhi	0.482	0.689	0.875

2.4.2 Genetic Distance Measure

Three distance measure were used to infer genetic distances between the populations under study. (1) D_A (Nei's genetic distance) is formulated for Infinite Allele Model (IAM) in which there is a rate of neutral mutation and each mutation give rise to a distinguishable allele (Nei, Tajima, & Tateno, 1983). D_A is calculated as

$$D_A = 1 - 1/r \sum_j^r \sum_i^{mj} \sqrt{X_{ij} Y_{ij}} \quad (7)$$

where X_{ij} and Y_{ij} are the frequencies of the *ith* allele at the *jth* locus in populations *X* and *Y*, respectively, and *mj* is the number of alleles at the *jth* locus. D_A was computed through a statistical program *Poptree*. (2) D_{ST} is the Nei' s standard genetic distance (Nei, 1972) given by

$$D_{ST} = -In\, J_{XY} / \sqrt{J_X\ J_Y} \quad (8)$$

where $J_X = \sum_j^r \sum_i^{mj} x_{ij}^{\ 2} / r$ and $J_Y = \sum_j^r \sum_i^{mj} y_{ij}^{\ 2} / r$ are the average heterozygosities over the loci for populations *X* and *Y*, respectively, and $J_{XY} = \sum_j^r \sum_i^{mj} x_{ij}\, y_{ij} / r$. D_{ST} was computed through *Phylip version 3.68.* (3) D_C is the chord distance proposed by Cavalli Sforza and Edward (1967). D_C is defined by

$$D_C = 2/\pi\ r \sum_j^r [\ 2\ (\ 1 - \sum_i^{mj} \sqrt{X_{ij}\, Y_{ij}})]^{1/2} \quad (9)$$

D_C was computed through *Phylip version 3.68.*

2.4.3 Coefficient of Variation (*CV*) of D_A, D_{ST} and D_C

Nei's distance and Cavalli Sforza's distance measures are different estimators of the same quantity under the same model. Therefore a measurement of relative variance (*CV*) was used for each distance measure. *CV* was calculated across 100 replicates of the original dataset obtained through re-sampling of the original dataset for each of the three distance measures. A random sample of fourteen populations (14 X 14 populations distance matrix) was used for this purpose.

2.4.4 Construction of Phylogenetic Trees

Three methods were used to reconstruct phylogenetic trees. (1) Neighbor Joining (NJ) Method. NJ method constructs a tree by successive clustering of lineages, setting branch lengths as the lineages join (Saitou, & Nei, 1987). (2) Unweighted Pair Group Method using Arithmetic Mean (UPGMA). UPGMA merge closest pair of taxa (by distance) and then recomputes distances to merged nodes via arithmetic mean of pairwise distances to leaves of the tree. (3) Continuous Character Maximum Likelihood Method (CONTML). This is a program in PHYLIP which estimates phylogenies by the restricted maximum likelihood method based on the Brownian motion model. It assumes that each locus evolves independently by pure genetic drift.

2.4.5 Consensus Tree

Consensus trees were generated by bootstrapping (100 to 1000 replications) of the original data taken from http://dnaa.bravehost.com/index.html and http://www.cstl.nist.gov/strbase/population/Omnipop (Figure 1 and Figure 2). Consensus trees were also constructed between the populations who either have a strong ethnic (Figure 3) or linguistic affiliation (Figure 4). Allele frequency data of Pakistani subpopulations were incorporated into the 62 world populations' data (*http://dnaa.bravehost.com/index.html*) as well as 36 world populations' data from Omnipop file (http://www.cstl.nist.gov/strbase/population/Omnipop). Two measures of genetic distance were used i.e. Nei's genetic distance (Figure 5 and Figure 6) and Cavalli Sforza chord distance (Figure 7 and Figure 8). Trees were constructed using NJ method.

2.4.6 Comparison with other Phylogenetic Trees

Consensus tree was then compared with the trees obtained from other molecular markers such as Alu and RFLP (Nei & Roychoudhury, 1993; Nei & Takezaki, 1996).

2.4.7 Principal Component Analysis (PCA)

PCA was performed for the allele frequency data of 62 world populations (http://dnaa.bravehost.com/index.html) (Figure 9). Pakistani subpopulation allele frequency data was incorporated into 62 world population data and PCA was performed again (Figure 10).

3. Results

- Coefficient of variation (*CV*) was in order of $D_{ST} > D_A > D_C$. Therefore, a consensus phylogenetic tree based on DC was constructed using NJ, UPGMA and Maximum Likelihood (ML) methods. NJ method showed higher bootstrap values. Comparison of the resultant trees showed that the tree topology was consistent with the trees reported for other molecular markers (Nei & Roychoudhury, 1993; Nei & Takezaki, 1996). Geographic, ethnic and linguistic demarcation between the populations was appreciable (Figure 1 through Figure 4).
- Tree topology was consistent with 'out of Africa' theory of human origin. African populations formed a distinct cluster with high bootstrap value (>950) and the remaining populations branched off from the African cluster.
- Geographical and ethnic demarcations between the populations were more obvious than linguistic demarcation i.e. populations who are in close geographical proximity to each other or who have a common ethnic origin showed tendency to form a separate cluster. For example, Chinese, Tibet, Bhutan, Thai, Malays and Japanese formed a separate cluster though they have diverse linguistic affiliations. China, Thailand and Malaysia belong to Austronesian class of languages while Bhutan, Tibet and Nepal belong to Sino Tibetan class. All these populations belong to the Mongoloid ethnic group. It showed that the STRs were more successful in delineating ethnic rather than linguistic partitioning.
- Phylogenetic efficiency of the three STRs for the populations and subpopulations of the Indian subcontinent was remarkable. Central Indian, south Indian and eastern Indian populations were well differentiated according to their ethnic and linguistic backgrounds. All the Dravidian speaking Australoid Golla subpopulations of Andhra Pradesh (seven in number) consistently formed a separate cluster in phylogenetic trees. Similarly, Eastern India castes Brahmin and Kayasth consistently showed a single cluster while another Eastern India caste Garo was closer to European Caucasians rather than their neighboring populations. Likewise Tamil Bohra muslims did not cluster with their neighboring Tamil sunni muslims, instead they were closer to the Mongoloid populations.
- PCA showed a distinct position of all the African populations in the of score plot of PC1 and PC2 (Figure 9). Indian populations and subpopulations were lying in the left upper quadrant while Mongoloids were in the right lower quadrant. Caucasoids were dispersed in the right half around the median axis.
- Phylogenetic tree (Figure 5 through Figure 8) showed a distinct cluster of the five Pakistani subpopulations with high bootstrap values ($\geq$ 840). It also showed the close affinity of Pakistani subpopulations to the Caucasoid and Mongoloid populations.
- PCA showed all the Pakistani subpopulations in the left lower quadrant except Muhajir that was in the left upper quadrant (Figure 10).

Figure 1. Phylogenetic tree of 62 world populations (*http://dnaa.bravehost.com/index.html*) based on Cavalli Sforza's Chord distance and NJ method (consensus tree= 1000 bootstrapping)

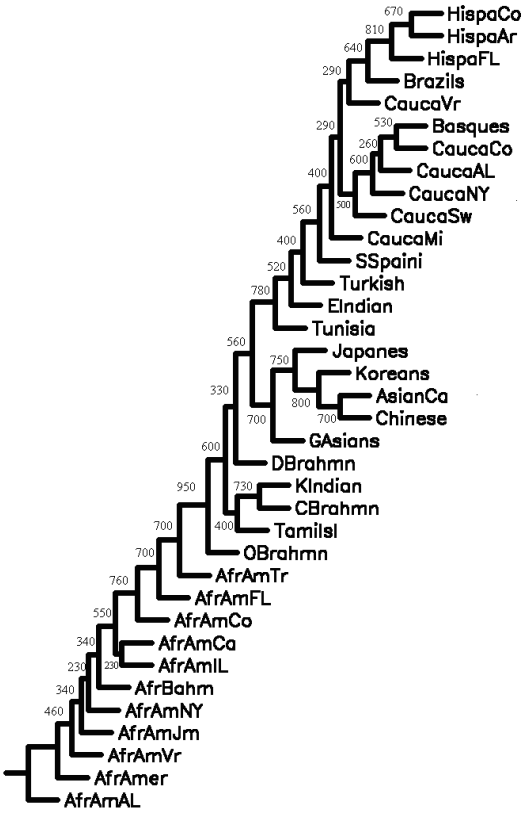

Figure 2. Phylogenetic tree of 36 world populations from Omnipop file (http://www.cstl.nist.gov/strbase/population/Omnipop) based on Cavalli Sforza's Chord distance and NJ method (consensus tree= 1000 bootstrapping)

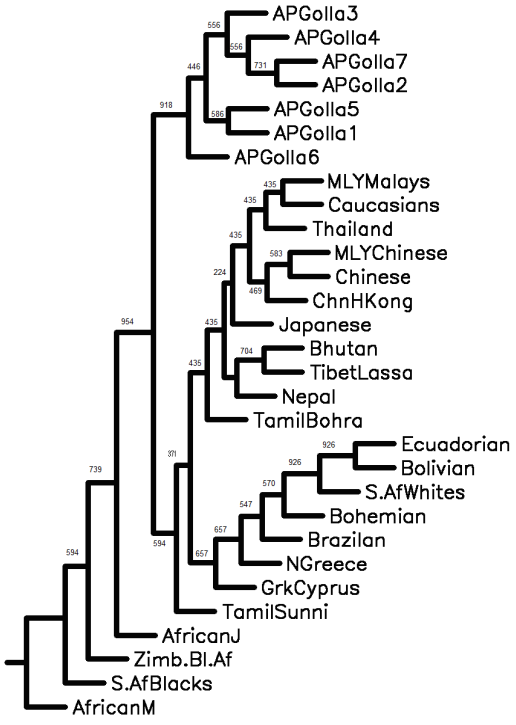

Figure 3. Phylogenetic tree of 30 world populations (*http://dnaa.bravehost.com/index.html*) with strong ethnic affiliations and less degree of genetic admixture based on Cavalli Sforza's Chord distance and NJ method (consensus tree = 1000 bootstrapping)

Figure 4. Phylogenetic tree of 35 world populations(*http://dnaa.bravehost.com/index.html*) with strong linguistic affiliations and less degree of genetic admixture based on Cavalli Sforza's Chord distance and NJ method (consensus tree = 1000 bootstrapping)

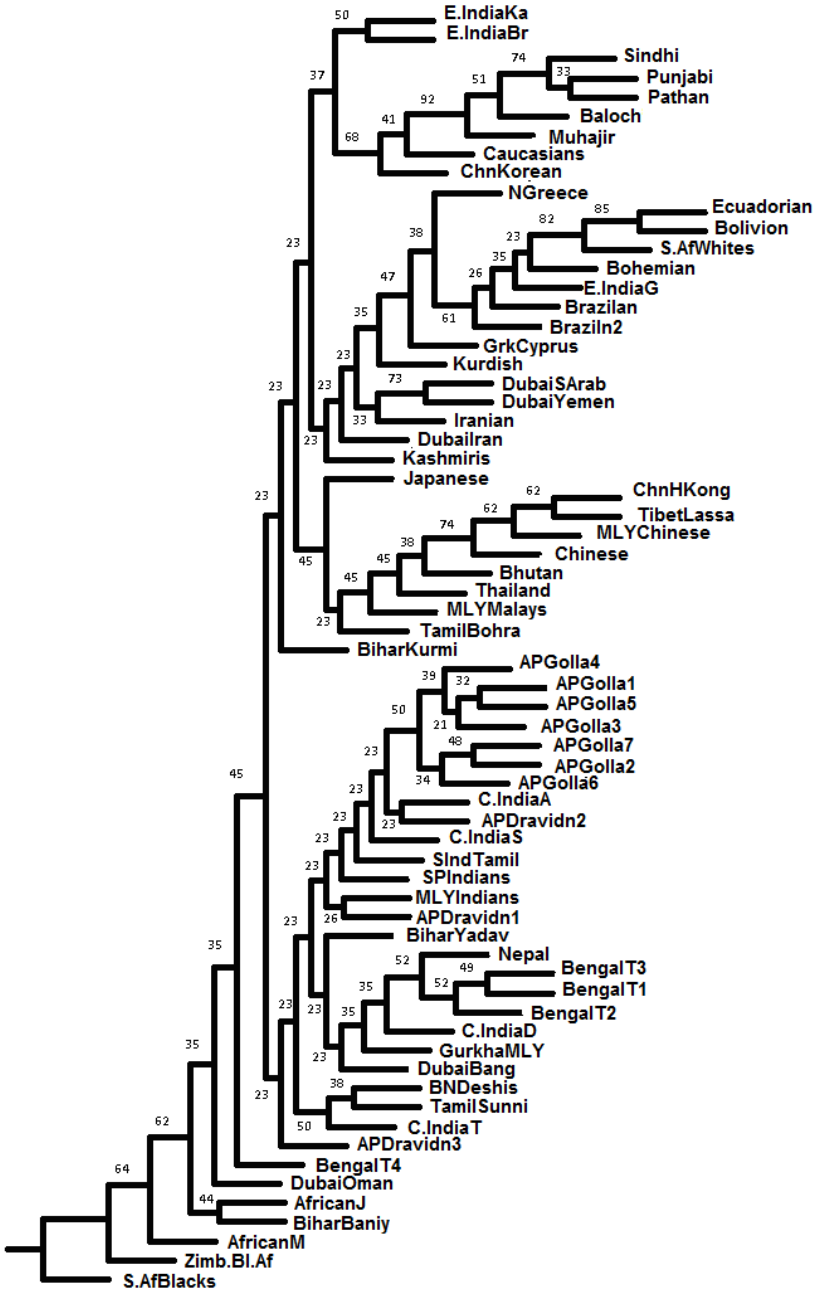

Figure 5. Phylogenetic tree of 67 world populations (*http://dnaa.bravehost.com/index.html*) including five Pakistani subpopulations of the present study. The phylogenetic tree is based on Nei's genetic distance and NJ method (consensus tree= 100 bootstrapping)

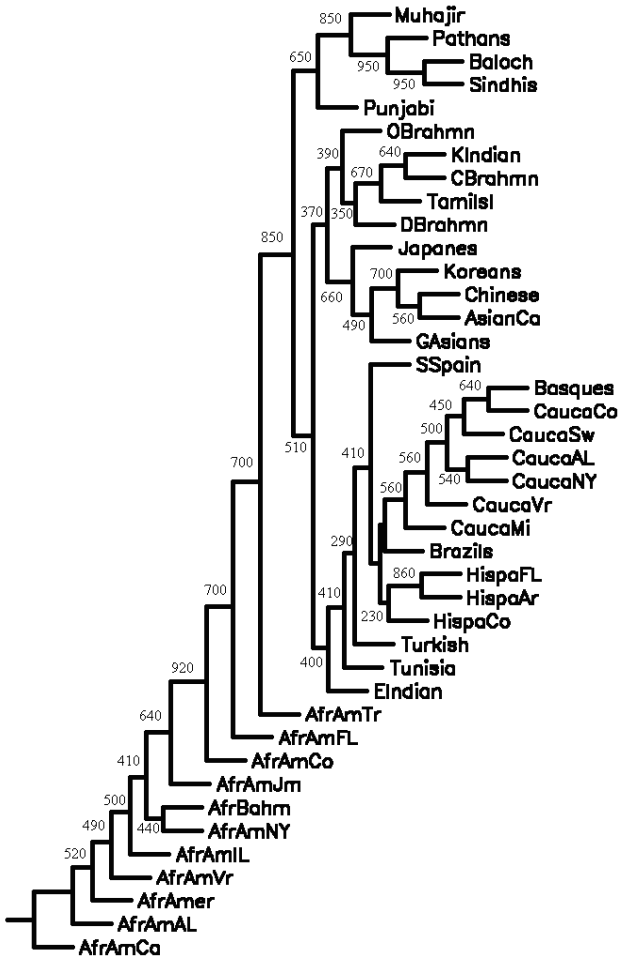

Figure 6. Phylogenetic tree of 41 world populations (*http://www.cstl.nist.gov/strbase/population/*Omnipop) including five Pakistani subpopulations of the present study. The phylogenetic tree is based on Nei's genetic distance and NJ method (consensus tree= 1000 bootstrapping)

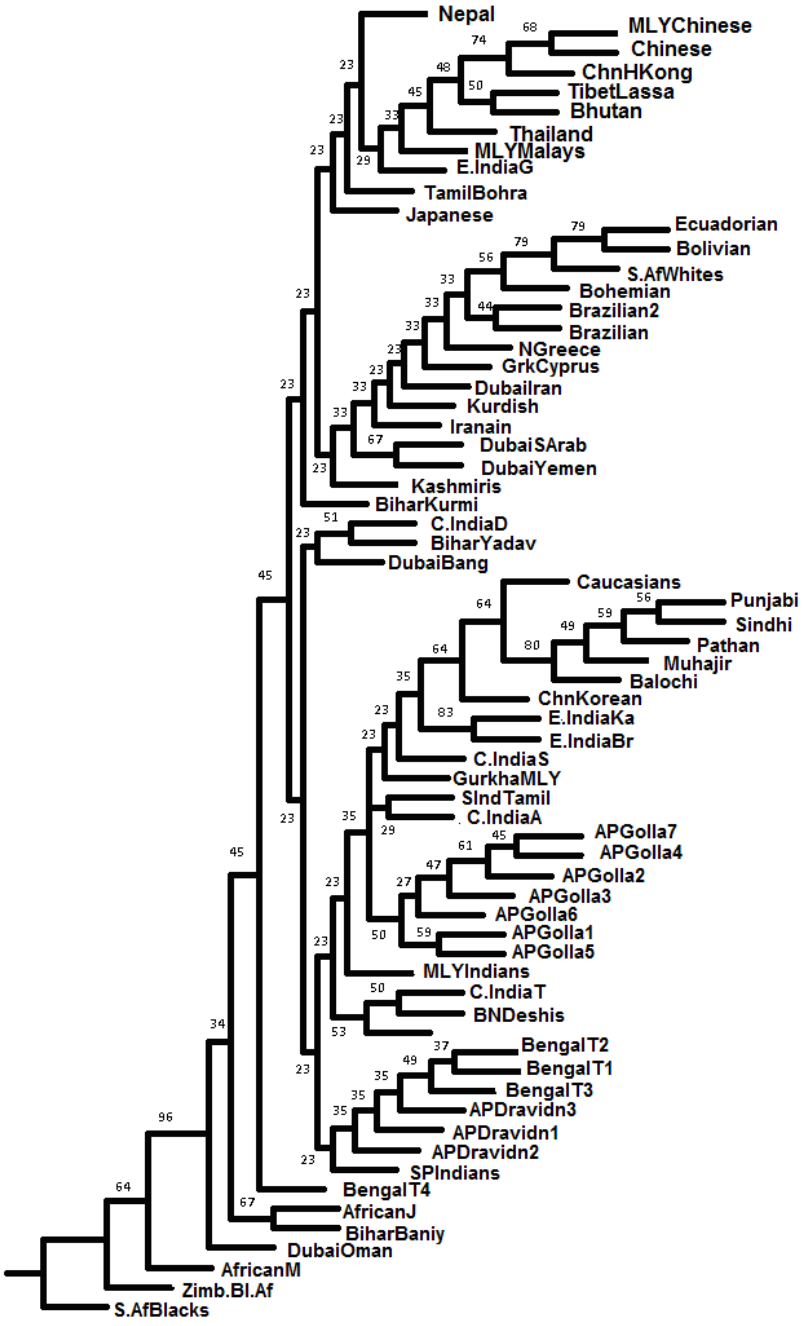

Figure 7. Phylogenetic tree of 67 world populations (*http://dnaa.bravehost.com/index.html*) including five Pakistani subpopulations of the present study. The phylogenetic tree is based on Cavalli Sforza's Chord distance and NJ method (consensus tree= 100 bootstrapping)

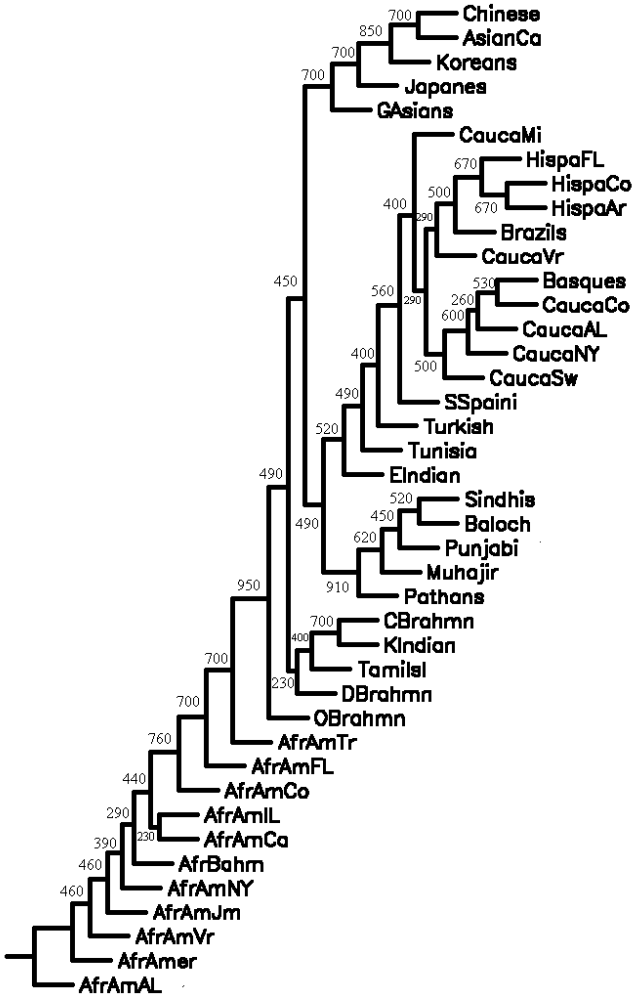

Figure 8. Phylogenetic tree of 41 world populations (*http://www.cstl.nist.gov/strbase/population/*Omnipop) including five Pakistani subpopulations of the present study. The phylogenetic tree is based on Cavalli Sforza's Chord distance and NJ method (consensus tree= 1000 bootstrapping)

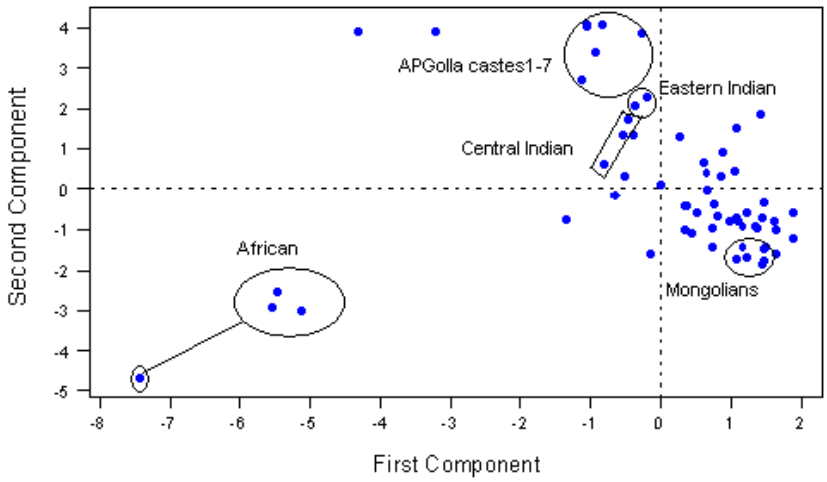

Figure 9. Score plot of PCA of 62 populations (*http://dnaa.bravehost.com/index.html*) based on allele frequencies of three STRs, CSF1PO, TPOX, and TH01

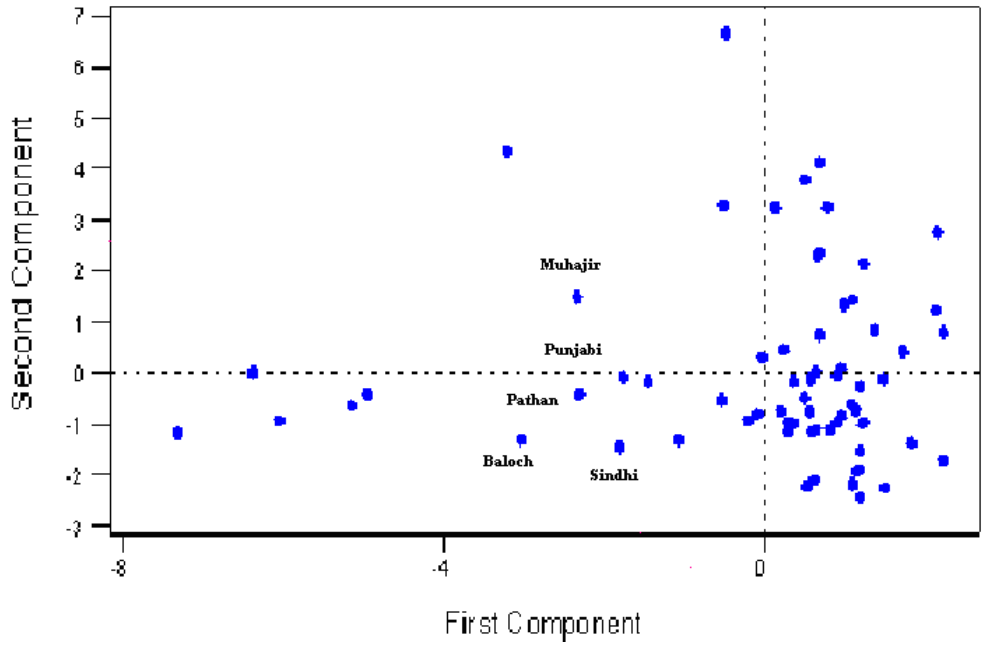

Figure 10. Score plot of PCA of 67 world populations (*http://dnaa.bravehost.com/index.html*) including five Pakistani subpopulations of the present study based on allele frequencies of three STRs, CSF1PO, TPOX, and TH01

4. Discussion

Evolutionary histories and phylogenetic relationship of many extant human populations have been explored using microsatellite loci (Bowcock et al., 1994; Deka et al., 1995; Gonser, Donnelly, Nicholson, & Rienzo, 2000; Rowold & Herrera, 2003). Rowold and Herrera (2003) and Agrawal and Faisal (2005) used five STR loci including CSF1PO, TPOX and TH01 for phylogenetic analyses and concluded that these STR loci are successful in reconstructing recent human evolutionary histories. However, they used only ten (10) and twenty one (21) population groups respectively in comparison to ninety eight (98) world populations used in the present study. Moreover the strategy employed in the present study was more comprehensive and each decision making step was explained logically. Findings were also supported by the validation studies.

Understanding the pattern and rate of mutations is very relevant to the applications of these hypervariable genetic markers in evolutionary studies as well as in gene mapping studies (Goldstein, Linares, Cavalli-Sforza, & Feldman, 1995; Shriver et al., 1995). The utility of a genetic marker for determining phylogenetic relationships within a given population is a function of the mutation rate of the marker and the overall genetic diversity of the examined population (Keim et al., 2004). When population genetic diversity is high, only markers with low mutation rates will yield accurate phylogenetic patterns. TPOX and TH01 showed the lowest mutation rates among all thirteen CODIS STR loci; hence likely to be suitable for phylogenetic purposes.

Topology of the phylogenetic trees (Figure 1 through Figure 4) was consistent with those obtained from other molecular markers. For example a phylogenetic tree for 26 human populations based on D_A using 29 polymorphic loci (Nei & Roychoudhury, 1993) showed the same partitioning of human populations as shown by the trees reconstructed in the present study. The trees were also compared with the trees based on RFLP data and Alu insertion polymorphism data using D_A distance measure (Nei & Takezaki, 1996). Tree topology and partitioning of the populations into ethnic groups were consistent with those of RFLP and Alu insertions. It should be mentioned that the performance of D_A distance measure in obtaining the correct tree topology is considered to be the same as that of D_C (Takezaki & Nei, 1996). Major ethnic groups identified in the present study were more or less similar to those recognized by classical anthropologist (Nei & Rouchaudhry, 1993). They were four in number namely, Negroid (Africans), Caucasoids (European and their related populations), Mongoloids (East Asians) and Australoid (Andhra Pradesh Golla castes). Tree topology was also supportive of 'out of Africa theory' which has gained popularity among geneticist and anthropologist during the last two decades (For example Nei, 1995; Templeton, 2002; Adams, 2008; Hanihara, 2008; Sun, Mullikin, Patterson, & Reich, 2009).

Stability of tree topology and the adequacy of the data to validate the topology are assessed by bootstrap values (Berry & Gascuel, 1996). Tree topology showed higher bootstrap values when applied to the dataset of populations who have lesser degree of admixture and a strong affiliation with a single ethnic or linguistic group (Figure 3 and Figure 4). It indicates that apart from the number of markers used, there are certain other factors which affect the phylogenetic efficiency of a marker. These factors include STR locus polymorphism, distance measures and methods to reconstruct phylogenetic trees (Nei & Roychoudhury, 1974; Nei, Kumar, & Takahashi, 1998; Goldstein & Pollock, 1994; Tajima & Takezaki, 1994; Takezaki & Nei, 1996; Takezaki & Nei, 2008). Ethnic demarcation showed more statistical support than linguistic demarcation. Another study using 182 autosomal microsatellites could not reveal any phylogenetic relationship between the two language isolate populations namely Hunza Burusho and Basques (Ayub et al., 2003). It was argued then that the microsatellites are best suited for the study of more recent population separations.

Phylogenetic efficiency of the three STR markers was worth noticing for the populations and subpopulations of the Indian subcontinent. Eastern India castes Brahmin and Kayasth consistently formed a single cluster. These two population groups are considered upper classes of Hindu caste system where intermarriages are not prohibited, while Garo which was closer to European Caucasians, is the middle class of Hindu caste system. Likewise Tamil Bohra Muslims were closer to the Mongoloid populations rather than their neighboring Tamil sunni Muslims. Bohra and Sunni are the two religious sects of Muslims between which marriages are generally prohibited. Dravidian speaking Golla castes of Andhra Pradesh remained separated from their neighboring subpopulations and branched off as a single cluster. Even within the Golla castes, western Golla castes (APGolla1 and APGolla5) were closer to each other than other Golla castes. The results established the efficiency of the three STRs (CSF1PO, TPOX, TH01) in delineating genetic relationships of the subpopulations of Indian subcontinent. The finding was validated for the subpopulations of Pakistan which is geographically a part of Indian subcontinent. All five Pakistani subpopulations namely Baloch, Muhajir, Pathan, Punjabi and Sindhi were united in a single cluster with a high bootstrap value that possibly suggests their common origin. Most of the Indian subcontinent populations are thought to be Caucasoid in origin (Cavalli Sforza, Menozzi, & Piazza, 1994). Pakistani subpopulations also showed their affiliation with other Caucasoid and Mongoloid populations. According to a hypothesis of populations' evolution and migration Indian subcontinent has been invaded by both the Caucasoid as well as Mongoloid populations (Nei & Roychoudhury, 1993). Mongoloids are also believed to originate from later splitting in Caucasoid race. Theories of gene flow and varying degrees of admixture between south Asian Indian populations and Mongoloid populations have also been proposed (Bamshad et al., 2003; Watkins, 2003; Shriver et al., 2005). Close affiliation of Pakistani subpopulations may be the results of gene admixture between the two populations. Another study based on HLA-A, -B, -C and –DRB, -DQB1 loci have also shown admixture of Pakistani ethnic groups with Caucasoids and Oriental populations (Mohyuddin, 2000).

In April 2011 American FBI recommended to remove few STR loci from the CODIS list due to their lower polymorphism observed across the world populations (Butler & Hill, 2012; Hares, 2012). TPOX was the least polymorphic of all the CODIS STR loci. In the present study TPOX showed heterozygosity values higher than those for CSF1PO (Table 5). Tri allelic pattern frequently observed for TPOX (Butler, 2005, Lane, 2008; Diaz, Rivas, & Carracedo, 2009) was not observed across the five subpopulations of Pakistan. Results emphasized that the STR loci should be investigated extensively for their efficiency as human identification markers across the populations and subpopulations of the Indian subcontinent. Heterogeneity of extant populations of the Indian subcontinent may deserve a separate standard set of STR loci. It is worth mentioning that European standard set of STR loci are/is different from American core set of STR loci (Butler & Hill, 2012).

It can be concluded that the three STRs successfully exhibited ethnic and linguistic as well as the(omit it) intra-ethnic differentiation across the populations and the subpopulations. Results also suggest that minimum number of markers can be used for reconstructing phylogenetic trees with high bootstrap values provided the markers are efficient for this purpose and a correct statistical strategy is employed. This study may help to identify the STR loci that can be used for forensic as well as phylogenetic purposes.

References

Adams, J. (2008). Human Evolutionary Tree. *Nature Education, 1(1)*. Retrieved from http://www.nature.com/scitable/topicpage/human-evolutionary-tree-417

Agrawal, S., & Khan, F. (2005). Reconstructing recent human phylogenies with forensic STR loci: A statistical approach. *BMC Genetics, 6,* 47. Retrieved from http://www.biomedcentral.com/1471-2156/6/47

Astolfi, P., Kidd, K. K., & Cavalli Sforza, L. L. (1981). A comparison of methods for reconstructing evolutionary trees. *Syst . Zool., 30,* 156-169. Retrieved from http://www.jstor.org/discover/10.2307/2992414

Ayub, Q., Mansoor, A., Ismail, M., Khaliq, S., Mohyuddin, A., Hameed, A., … & Mehdi, S. Q. (2003). Reconstruction of human evolutionary tree using polymorphic autosomal microsatellites. *Am. J. Phys. Anthrop., 122*, 259-268. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/14533184

Bamshad, M. J., Wooding, S., Watkins, W. S., Ostler, C. T., Batzer, M. A., & Jorde, L. B. (2003). Human population genetic structure and inference of group membership. *Am. J. Hum. Genet., 75,* 578-589. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1180234/

Berry, V., & Gascuel, O. (1996). On the interpretation of bootstrap trees: Appropriate threshold of clade selection and induced gain. *Mol. Biol. Evol., 13,* 999-1011. http://mbe.oxfordjournals.org/cgi/doi/10.1093/molbev/13.7.999

Bowcock, A. M., Ruiz – Linares, A., Tomfohrde, J., Minch, E., Kidd, J. R., & Cavalli Sforza, L. L. (1994). High resolution of human evolutionary trees with polymorphic microsatellites. *Nature, 368,* 455-457. Retrieved from http://www.nature.com/nature/journal/v368/n6470/abs/368455a0.html

Budowle, B., Moretti, T. R., Baumstark, A. L., Defenbaugh, D. A., & Keys, K. M. (1999). Population data on the thirteen CODIS core short tandem repeat loci in African Americans, U.S. Caucasians, Hispanics, Bahmians, Jamaicans and Trinidadians. *J. Forensic. Sci., 44*, 1277-1286. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/ 10582369

Budowle, B., Moretti, T. R., Niezgoda, N. R., & Brown, B. L. (1998). CODIS and PCR based short tandem repeat loci: Law enforcement tools. *Second European symposium on human identification. Promega corporation,* 73-88.

Butler, J. M. (2005). *Forensic DNA typing: Biology, technology and genetics of STR markers*. USA: Elsevier.

Butler, J. M. (2006). Genetics and genomics of core short tandem repeat loci used in human identity testing. *J. Forensic. Sci., 51,* 253-265. http://www.cstl.nist.gov/strbase/pub_pres/Butler2006JFS

Butler, J. M., & Hill, C. R. (2012). Biology and genetics of new autosomal STR loci useful for forensic DNA analysis. *Forensic Sci. Rev., 24,* 15-26. Retrieved from http://www.cstl.nist.gov/strbase/pub_pres/Butler-Hill-FSR2012-newSTRloci.

Cavalli Sforza, L. L., & Edwards, W. (1967). Phylogenetic analysis: Models and estimation procedures. *Am. J. Hum. Genet., 19,* 233-257. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1706274/

Cavalli Sforza, L. L., Menozzi, P., & Piazza, A. (1994). *The history and geography of human genes*. New Jersey: Princeton University Press.

Deka, R., Jin, L., Shriver, M. D., Yu, L. M., Decroo, S., Hundrieser, J., … & Chakraborty, R. (1995). Population genetics of dinucleotide (dCdA);(dG-dT), polymorphisms in world populations. *Am. J. Hum. Genet., 56,* 461-474. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1801145/

Diaz, V., Rivas, P., & Carracedo, A. (2009). The presence of tri-allelic TPOX genotypes in Dominican Population. *Forensic Sci. Int. Genetics Supplement Series., 2,* 371-372. http://dx.doi.org/10.1016/j.fsigss. 2009.09.021

Felsenstein, J. (2003). *Inferring phylogenies.* Sunderland, MA: Sinauer Associates.

Goldstein, D., Linares, A., Cavalli-Sforza, L., & Feldman, M. (1995). An evaluation of genetic distances for use with microsatellite loci. *Genetics, 139*, 463–471. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1206344/

Goldstein, D. B., & Pollock, D. D. (1994). Least square estimation of molecular distance- noise abatement in phylogenetic reconstruction. *Theor. Popul. Biol., 44,* 219-226. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8066551

Gonser, R., Donnelly, P., Nicholson, G., & Rienzo, A. D. (2000). Microsatellite mutations and inferences about human demography. *Genetics, 154,* 1793-1807. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1461043

Hanihara, T. (2008). Morphological variation of major human populations based on nonmetric dental traits. *Am. J. Phys. Anthrop, 136,* 169-182. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/ajpa.20792/abstract

Hares, D. R. (2012). Expanding the CODIS core loci in United States. *Forensic Sci. Int. Genetics, 6*, e52-e54. http://dx.doi.org/10.1016/j.fsigen.2011.04.012

Hedrick, P. W. (2011). *Genetics of populations* (4th ed.). Sudbury, Massachusetts: Jones and Bartlett publishers.

Holder, M., & Lewis, P. O. (2003). Phylogeny estimation: Traditional and Bayesian approaches. *Nature Genetics, 4,* 275-284. Retrieved from http://www.nature.com/nrg/journal/v4/n4/full/nrg1044.html

Keim, P., Matthew, N. V. E., Talima, P., Amy, J. V., Lynn, Y. H., & David, M. W. (2004). Anthrax molecular epidemiology and forensics: Using the appropriate marker for different evolutionary scales. *Infection, Genetics and Evolution., 4,* 205–213. Retrieved from http://jan.ucc.nau.edu/aa238/Kenefic%20reference1

Kobilinsky, L., Liotti, T. F., & Oeser Sweat, J. (2005). *DNA: Forensic and legal applications*. Hoboken, New Jersey: Wiley International.

Lane, A. B. (2008). The nature of tri allelic TPOX genotypes in African populations. *Forensic Sci. Int. Genetics, 2,* 134-137. http://www.ncbi.nlm.nih.gov/pubmed/19083808.

Li, C. C. (1976). *First course in population genetics.* Boxwoods. Pacific Grove, CA. USA.

Li, S., Pearl, D. K., & Doss, H. (2000). Phylogenetic tree construction using Markov chain Monte Carlo. *J. Am. Stat. Assoc., 95*, 493-508. Retrieved from http://www.stat.ufl.edu/~doss/Research/mc-trees.pdf

Mohyuddin, A. (2000). *Genetic diversity of Pakistani subpopulations*. PhD Dissertation, Quaid-e-Azam University, Islamabad. Retrieved from http://eprints.hec.gov.pk/2110/1/2028.htm

Nei, M. (1972). Genetic distance between populations. *The American Naturalist., 106,* 283-292. Retrieved from http://www.jstor.org/discover/10.2307/2459777.

Nei, M. (1995). Genetic support for the out of Africa theory of human evolution. *Proc. Natl. Acad. Sci., 92,* 6720-6722. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC41400/

Nei, M., & Roychoudhury, A. K. (1974). Genic variation within and between the three major races of man, caucasoids, negroids, and mongoloids. *Am. J. Hum. Genet., 26,* 421-443. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1762596/

Nei, M., & Roychoudhury, A. K. (1993). Evolutionary relationship of human populations on global scale. *Mol. Biol. Evol., 10,* 927-943. Retrieved from http://mbe.oxfordjournals.org/content/10/5/927

Nei, M., & Takezaki, N. (1996). The root of the phylogenetic tree of human populations. *Mol. Biol. Evol., 13,* 170-177. http://mbe.oxfordjournals.org/content/13/1/170.

Nei, M., Kumar, S., & Takahashi, K. (1998). The optimization principle in phylogenetic analysis tends to give incorrect topologies when the number of nucleotides or amino acids used is small. *Proc. Natl. Acad. Sci. USA, 95,* 1239-12397. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9770497

Nei, M., Tajima, F., & Tateno, Y. (1983). Accuracy of estimated phylogenetic trees from molecular data, II gene frequency data. *J. Mol. Evol., 19,* 153-170. Retrieved from http://link.springer.com/article/10.1007%2FBF02300753

Rowold, D. J., & Herrera, R. J. (2003). Inferring recent human phylogenies using forensic STR technology. *Forensic Sci. Int., 133*, 260-265. Retrieved from http://www.fsijournal.org/article/S0379-0738(03)00073-2/fulltext

Saitou, N., & Nei, M. (1987). The neighbor joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol., 4,* 406-425. Retrieved from http://mbe.oxfordjournals.org/content/4/4/406

Shete, S., Tiwari, H., & Elston, R. C. (2000). On estimating the heterozygosity and polymorphism information content value. *Theor. Popul. Biol., 57,* 265-271. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10828218

Shriver, M., Jin, L., Boerwinkle, E., Deka, R., Ferrell, E., & Chakraborty, R. (1995). A novel measure of genetic distance for highly polymorphic tandem repeat loci. *Mol. Biol. Evol., 12,* 914–920. Retrieved from http://mbe.oxfordjournals.org/content/12/5/914.short

Shriver, M., Mei, R., Parra, E. J., Sonpar, V., Halder, I., Tishkoff, S. A., … Jones, K. W. (2005). Large scale SNP analysis reveals clustered and continuous patterns of human genetic variation. *Human Genomics, 2*, 81-89. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16004724.

Sun, J. X., Mullikin, J. C., Patterson, N., & Reich, D. E. (2009). Microsatellites are molecular clocks that support accurate inferences about history. *Mol. Biol. Evol., 26,* 1017-1027. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2734136/

Tajima, F., & Takezaki, N. (1994). Estimation of evolutionary distance for reconstructing molecular phylogenetic trees. *Mol. Biol. Evol., 11,* 278-286. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/ 8170368

Takezaki, N., & Nei, M. (1996). Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. *Genetics, 144,* 389-399. Retrieved from http://www.genetics.org/content/144/1/389

Takezaki, N., & Nei, M. (2008). Empirical tests of the reliability of phylogenetic trees constructed with microsatellite DNA. *Genetics, 178,* 385-392. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2206087/

Templeton, A. (2002). Out of Africa again and again. *Nature, 416,* 45-51. Retrieved from http://www.nature.com/nature/journal/v416/n6876/abs/416045a.html

Watkins, W. S., Rogers, A. R., Ostler, C. T., Wooding, S., Bamshad, M. J., Brassington, A. E., … Jorde, L. B. (2003). Genetic variation among world populations: Inferences from 100 Alu insertion polymorphisms. *Genome Research, 13*, 1607-1618. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12805277

Zhivotovsky, L. A., & Feldman, M. W. (1995). Microsatellite variability and genetic distances. *Proc. Natl. Acad. Sci. USA, 92*, 11549-11552. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC40439/