Journal of Positive School Psychology http://journalppw.com
2022, Vol. 6, No. 9, 3602-3612

Precision Id Mtdna Panel Examines Maternal Lineages And Ancient Migrations In Indian Groups And Identifies Region-Specific Mitochondrial Haplotypes

Swati Arora[1]*, Ajay Kumar[2], Rajiv Kumar[3],

*[1]*Department of Biosciences, School of Basic and Applied Sciences,Galgotias University, Greater Noida.*
[2]*Associate Professor, Department of Biosciences, School of Basic andApplied Sciences, Galgotias University, Greater Noida.*
[3]*Associate Professor, Department of Biosciences, School of Basic andApplied Sciences, Galgotias University, Greater Noida.*

*Corresponding Author: Swati Arora

Abstract

Introduction: Uniparentally inherited molecular markers for forensics and other applications where human identification from tainted samples is a crucial problem. The current study aimedto identify region-specific mitochondrial haplotypes for human identification from complex samples by assessing the Precision ID mtDNA panels, examining maternal lineages, and investigating prehistoric migration episodes in Indian populations.
Materials and Methods: We considered 40 unrelated people from India's Eastern, Northern, Western, and Southern regions when processing the extracted DNA samples to resemble forensicspecimens. In addition, 38 Indian sequences had already been for the haplotype-based evaluationof mtDNA-based markers.
Results: Among those investigated, the mitochondrial macro-haplogroup M predominated (58%). Due to their distinct coalescent histories, we projected various expansion dates for NorthIndians (26kya), East Indians (22kya), and West Indians (15kya). However, because of frequentfree mingling and the quick expansion of the Indo-European language, these populations are admixed and lack any meaningful subpopulation structure. Due to the high incidence of endogamy in this area, we found a substantially older expansion time (28kya) and minimal genetic variation among South Indians. Finally, we have discovered seven hotspot sites relevantfor human identification: five West Indian-specific (16069, 16169, 16206, 215 & 243), four North Indian-specific (16170, 16181, 16185 & 285), three East Indian-specific (16224, 16344 &41), and one South Indian-specific (480). To validate the results of this pilot-scale study, however, a more in-depth investigation with a bigger cohort and a variety of genetic markers is required.
Conclusion: The Precision ID CR panel for human identification by this comprehensivepilot-scale genetic investigation evaluates the Precision ID mtDNA panels on unrelated individuals from four zones of India.

Keywords: Precision ID mtDNA Panel; region-specific mitochondrial haplotypes; hotspot mtDNA positions; mitochondrial haplogroup diversity; human identification from challenging samples.

Introduction

In the current Next-Generation Sequencing (NGS) era, the two widely used molecular marker types for human identification are DNA sequence polymorphisms and DNA repeat variations. Itis due to the ability of the latest advanced NGS platforms to generate unbiased high-throughput data cost-effectively, as anticipated in an earlier study (1). The maternally inherited mitochondrial DNA (mtDNA) and the paternally inherited Y-chromosome are reasonably effective in human identification through inferring ancestry information. Due to a lack of recombination, polymorphisms in these portions of the genome can provide ancestry identification by determining variable haplotypes and corresponding haplogroups (HGs).

Moreover, we believe that several successive pre-historic migration events caused the evolutionof these HGs (2). Therefore, studying their patterns is also essential for human identification by reconstructing ancestral genetic history (3).

Among these two markers, mtDNA-based markers are advantageous in forensic, medical, and population genetics studies (1, 4). Firstly, its mutation rate is more rapid than the nuclear genome(5), especially the control region (CR) or the displacement loop (D-loop) region, which evolves approximately five times more rapidly than the rest of the mitogenome (mt-genome) (6). This D-loop segment is the most polymorphic region, and it has two hypervariable regions,

Hypervariable Region I (HVR-I), which spans around 16024-16569bp, and Hypervariable Region II (HVR-II), which spans around 1-576bp (7). Secondly, being the powerhouse of the cell, mitochondria, bearing the circular double-stranded DNA, are present in hundreds to thousands of copies, whereas nuclear DNA only has two copies per diploid cell. It increases the sensitivity of mtDNA-based analysis and the probability of successful PCR amplification, even from very challenging biological or forensic specimens that contain a limited amount of DNA ordegraded DNA (8). Thirdly, being maternally inherited, there is no chance of recombination.

Still, it has a higher substitution rate and follows a non-Mendelian inheritance pattern (9). Consequently, it can effectively establish lineages with close or distant relationships and differentiate closely related individuals. Moreover, some earlier studies have shown that the hypervariable regions of the mt-genome can as an individual identification marker or barcode inhumans (10).

In the pre-NGS era, there was a limitation in sequencing the whole mt-genome using Sanger sequencing due to practical and technical reasons. However, due to modern high-throughput sequencing platforms, massively parallel sequencing has become a routine job, and we can perform whole-genome sequencing in a simplified manner (8). In this context, the recent development of the Applied Biosystems™ Precision ID System facilitates a better prospect of generating whole mt-genome (16569bp) sequencing data from highly compromised samples (4, 11). The system contains Precision ID mtDNA Panels, a solution for automated library and template preparation using the Ion Chef™ system, sequencing and analysis with Ion S5™ Semiconductor Sequencer, and Converge™ software (12). The whole-genome panel uses two multiplexed primer pools with 81 primer pairs each to target the whole mt-genome. Furthermore, the control area panel uses two multiplexed primer pools with seven primer pairs each to target both hypervariable regions of the mt-genome.

The effectiveness of these Precision ID panels on diverse mainland Indian populations to identify hotspot mt-genome positions facilitating human identification from challenging samples is yet tobe ascertained. Therefore, the current study aimed to assess the Precision ID system on some degraded samples collected randomly from unrelated individuals to determine the mitochondrial HGs diversity and the probable migration patterns of studied population groups. We also attempted to identify some region-specific hotspot mitochondrial polymorphisms for aiding human identification in forensic and other applications. Moreover, we endeavored to provide an analytical workflow, from sample processing to data analysis, which could be effective in humanidentification through inferring ancestral genetic history from any molecular marker.

Materials and methods

Subject Details

We randomly collected buccal swab samples from unrelated individuals from the East, North, South, and West zones of India and finally considered ten individuals from each zone for this study who were age and ethnically-matched and provided duly signed informed consent. In addition, we filled up a questionnaire with essential demographic data like age, weight, gender,type and duration of smoking and drinking habits, and other related information from each participant through personal interviews. We carried out all the methods following relevant guidelines and regulations.

Extraction of genomic DNA and its modifications

The genomic DNA (gDNA) was extracted from the buccal swab samples using the PrepFiler Express™ Forensic DNA Extraction Kit (Cat. No 4441352) in the AutoMate Express™ Instrument (Applied Biosystems, USA) and quantified by Qubit® 2.0 Fluorometer (Thermo Fisher Scientific, USA). Then they were treated with 300 ng/µl of Humic acid (Sigma Aldrich, SKU: 53680) and subjected to random fragmentation through sonication in COVARIS (p/n 600028) for 20-30 minutes to mimic compromised forensic samples. Furthermore, we assessed the degree of fragmentation through the E-gel Electrophoresis system (Thermo Fisher Scientific,USA).

Assessment of Applied Biosystems™ Precision ID System

In this study, initially, we ran the Precision ID mtDNA Control Region (CR) Panel (Cat. No A31443) to get the mtDNA D-loop region information. Afterward, to verify CR panel data, weran the Precision ID mtDNA Whole Genome (WG) Panel (Cat. No. A30938). We prepared the library from the treated DNA samples on the Ion

Chef™ robotics system (12) and quantified the amplified pooled libraries in qPCR using the Quantifiler™ Trio DNA Quantification Kit (Cat. No 4482910). The final libraries were clonally amplified via emulsion PCR and loaded onto Ion S5™ 530 sequencing chip to carry out sequencing in the Ion S5™ instrument. After sequencing,we used Converge™ v2.3 software for primary analysis of raw sequencing data (from BAM to VCF) and exported the observed mt-genome variants into XLSX format for further analysis.

Phylogenetic analysis

We determined mtDNA HGs based on the complete mt-genome [1-16569] and control region [16024-16569; 1-576] using Haplogrep2 (13) and EMPOP Haplogroup Browser *(https://empop.online/hg_tree_browser)*. Then we constructed the Quasi-median (QM) networkusing the EMPOP-NETWORK tool (https://empop.online/network) employing the EMPOPspeedy filter. The generated HGs network was visualized and adjusted in the DrawNetWork v1.24 tool *(https://empop.online/downloads*).

In addition, we inferred the evolutionary history by employing the Maximum Likelihood (ML) method based on the Hasegawa-Kishino-Yano (HKY) model (14) with 1000 bootstrap replicates (15) in MEGA7 (16). Apart from our generated 40 sequences, we considered 38 other Mainland Indian (North, West, Central, and South Indian) sequences for the analysis (Table 1) and rooted the tree with one Chimpanzee sequence (Accession no. U84293.1) and 5 Neanderthals sequences (Accession no. AM948965.1, DQ836132.1, EU078680.1, FM865410.1, and KX198087.1), as used in a recent study (17).

Statistical approaches

We used ARLEQUIN v3.5.2.2 software (18) to estimate the molecular diversity indices, mean pairwise differences (MPD), nucleotide diversity (π), haplotype diversity (Hd), initial theta (θa),tau (τ), raggedness index (r), and the number of migrants (M). We also studied mismatch distribution to assess demographic dynamics under the spatial expansion model. We estimated the departure from neutrality by Fu's Fs and Tajima's D based on 1000 coalescent simulations. The genetic structure of the studied populations was evaluated by the Analyses of MOlecular VAriance (AMOVA) based on 10000 permutations in ARLEQUIN v3.1 software (19). Besides, Effective population size (Ne) and population expansion age (AYa) were also calculated as suggested in a recent study (17). Keeping because of our moderate sample size, we performed the R2 Test in DnaSP v5.10.01 software (20) based on 1000 coalescent simulations, as the behavior of this test was found superior for small sample sizes (21). Furthermore, we evaluated marginal likelihoods to estimate the migration flow among the study subjects and compared different migration models using MIGRATE-n v4.4.3 (22). Lastly, we generated the bi-directional migration plot (23) in R-Studio *(https://www.rstudio.com/)* to understand the probable courses of gene flows among the studied Indian populations.

Table 1. Information on 18 sample populations

Sl. No.	Region/State		Population	Number of samples	Language family	Reference
1	Mainland Indian	Eastern Region	Mixed	10	Indo-European	**Present Study**
2		Odisha	Paudibhuiya	5		PopSet ID: 116242229
3		Jharkhand	Munda	5	Austroasiatic/ Mon–Khmer	[1] Thangaraj et. al.; *Human genetics* (2005); 116: 507-517 [2] PopSet ID: 154814435
4			Santhal	6		Thangaraj et. al.; *Human genetics* (2005); 116: 507-517
5			Oraon	2	Dravidian	
6		Western Region	Mixed	10	Indo-European	**Present Study**
7		Northern Region	Mixed	10	Indo-European	**Present Study**
8		Uttar Pradesh	Harijan	2	Indo-European	Thangaraj et. al.; *Human genetics* (2005); 116: 507-517
9			Rajput	2		
10			Yadava	2		
11			Other	1		GenBank ID: AJ234978.1
12		Chhattisgarh	Kanwar	2	Indo-European	Thangaraj et. al.; *Human genetics* (2005); 116: 507-517
13		Madhya Pradesh	Bharia	2	Dravidian	Thangaraj et. al.; *Human genetics* (2005); 116: 507-517
14	South Indian	Southern Region	Mixed	10	Dravidian	**Present Study**
15		Andhra Pradesh	Yanadi	3		Thangaraj et. al.; *Human genetics* (2005); 116: 507-517
16		TamilNadu	Baduga	1		Thangaraj *et. al.*; *Human genetics* (2005); 116: 507-517
17			Oorali	3		
18	Northeast Indian	Assam	Sylheti	7	Indo-European	Kundu et. al.; *Gene* (2021); 813: 146098

Results

Demographic characteristics of the studied individuals

The summary statistic of the studied Indian subject demography, mentioned in **Table 2**, has shown that the mean age (±SE) of our studied individuals was 40.6 (±0.8) years and the mean body weight (±SE) of our studied individuals was 69.4 (±1.5) KG. Among the studied individuals, 65% were females. We observed a significant association between body weight and gender in our studied populations (p-value: 0.0001). We also observed that most females (61.5%) fell under the non-smoker and non-drinker group, but among the male subjects, 35.7% were moderate smokers, and 42.9% of them were moderate smokers as well as drinkers. The gender-wise distinction in non-dietary habits was statistically significant (p-value: 0.0058). Besides, we studied the combined effect of studied demographic parameters in our studied populations (Table 3), which showed that

combinations of bodyweight+non-dietary habits+gender and age+non-dietary habits+gender were significantly associated (p-value: 0.0173and 0.0111, respectively). Moreover, the variety of age+bodyweight+non-dietary habits+gender showed a strong association (p-value: 0.0468) in our studied individuals **(Supplementary Table S1)**.

Table 2. Subjects' demographics

Variables		No. (%) of Participants FEMALE (n=26)	MALE (n=14)	Pearson Chi-Square p-value
AGE				
	Mean ± SE	40.2 ± 1.1	41.4 ± 0.7	
	> 40.6 years	9 (34.6)	8 (57.1)	0.1692
	≤ 40.6 years	17 (65.4)	6 (42.9)	
BODYWEIGHT				
	Mean ± SE	65.0 ± 1.2	77.5 ± 2.4	
	> 69.4 KG	6 (23.1)	12 (85.7)	**0.0001**
	≤ 69.4 KG	20 (76.9)	2 (14.3)	
HABITS				
	Only smoker	4 (15.4)	5 (35.7)	
	No Smoking	4 (15.4)	-	**0.0058**
	Smoking + Drinking	2 (7.7)	6 (42.9)	
	No Smoking + No Drinking	16 (61.5)	3 (21.4)	
REGION				
	EAST	8 (30.8)	2 (14.3)	
	NORTH	4 (15.4)	6 (42.9)	0.1843
	SOUTH	8 (30.8)	2 (14.3)	
	WEST	6 (23.1)	4 (28.6)	

Note: P-value < 0.05 considered as statistically significant (bold). **Only smoker:** smoker 10-19 cigarettes per day; **Smoking +Drinking:** smoker 10-19 cigarettes per day and drinks 60-100 per day.

Sequencing results

In this study, after the treatment, the mean length of our gDNA fragments was ~125bp **(Supplementary Figure S1)**. For the CR panel, we achieved an average depth (±SE) of 3,240.7X (±837.6X) with 95.3% (±1.0%) mean (±SE) sequence uniformity. In contrast, for theWG panel, we achieved an average depth (±SE) of 5,565.3X (±779.9X) with the mean (±SE) sequence uniformity of 90.8% (±2.7%) using the Ion S5TM 530 sequencing chips.

Supplementary Table S2 describes the detailed summary statistics of our sequencing runs.

Status of the observed mitochondrial haplogroup diversity

Keeping views on the nature of the samples, we initially assessed the mtDNA D-loop region (16024-16569; 1-576) using the CR panel and determined the HGs based on it (Supplementary Figure S2). Afterward, we verified this with the representative sequencing of the whole mt-genome (1-16569) using the WG panel **(Supplementary Table S3)**. We observed that mitochondrial HGs affiliation from both methods was concordant **(Supplementary Table S4)**, as we observed a match in 86.4% of cases. However, mitochondrial macro-haplogroup (macro-HG) affiliation was matched absolutely (100%). Furthermore, it revealed that among our studied 40 individuals from four different zones of India, 57.5% (23/40) of them belonged to the non-African haplogroup M, followed by R (20%; 8/40), other descendants of R like H / J / U / T (17.5%; 7/40), and haplogroup W (5%; 2/40). Among the observed individuals bearing the M macro-HGs, 39.1% (9/23) of them belonged to the Eastern region, followed by Western (21.7%; 5/23), Southern (21.7%; 5/23), and Northern (17.4%; 4/23) region. In contrast, North Indians mostly carried the haplogroup R and its descendants (40.0%; 6/15), followed by Western (33.3%; 5/15), Southern (20.0%; 3/15), and Eastern (6.7%; 1/15) Indians. However, the population-wise prevalence of observed HGs showed East Indians bearing the M haplogroup (90%; 9/10) and North Indians bearing the R and its descendant haplogroups (60%; 6/10).

Table 3. Status of studied demographic parameters in dual combinations

Variables		No. (%) of Participants FEMALE (n=26)	MALE (n=14)	Pearson Chi-Square p-value
BODYWEIGHT	HABITS			
> 69.4 KG	Only smoker	1 (3.8)	5 (35.7)	
	Smoking + Drinking	1 (3.8)	4 (28.6)	
	No Smoking + No Drinking	4 (15.4)	3 (21.4)	
≤ 69.4 KG	Only smoker	3 (11.5)	-	**0.0173**
	No Smoking	4 (15.4)	-	
	Smoking + Drinking	1 (3.8)	2 (14.3)	
	No Smoking + No Drinking	12 (46.2)	-	
AGE	HABITS			
> 40.6 years	Only smoker	2 (7.7)	1 (7.1)	
	No Smoking	1 (3.8)	-	
	Smoking + Drinking	1 (3.8)	5 (35.7)	
	No Smoking + No Drinking	5 (19.2)	2 (14.3)	**0.0111**
≤ 40.6 years	Only smoker	2 (7.7)	4 (28.6)	
	No Smoking	3 (11.5)	-	
	Smoking + Drinking	1 (3.8)	1 (7.1)	
	No Smoking + No Drinking	11 (42.3)	1 (7.1)	
REGION	HABITS			
EAST	Only smoker	2 (7.7)	1 (7.1)	
	No Smoking	1 (3.8)	-	
	Smoking + Drinking	1 (3.8)	1 (7.1)	
	No Smoking + No Drinking	4 (15.4)	-	
NORTH	Only smoker	-	2 (14.3)	
	Smoking + Drinking	1 (3.8)	3 (21.4)	
	No Smoking + No Drinking	3 (11.5)	1 (7.1)	
SOUTH	Only smoker	1 (3.8)	1 (7.1)	0.5739
	No Smoking	2 (7.7)	-	
	Smoking + Drinking	-	1 (7.1)	
	No Smoking + No Drinking	5 (19.2)	-	
WEST	Only smoker	1 (3.8)	1 (7.1)	
	No Smoking	1 (3.8)	-	
	Smoking + Drinking	-	1 (7.1)	
	No Smoking + No Drinking	4 (15.4)	2 (14.3)	

Note: P-value < 0.05 considered as statistically significant (bold). **Only smoker:** smoker 10-19 cigarettes per day; **Smoking +Drinking:** smoker 10-19 cigarettes per day and drinks 60-100 per day.

Determining region-specific haplotypes

Upon analyzing mitochondrial HV1 and HV2, we observed 26 haplotypes, of which eight haplotypes (30.8%) were shared **(Table 4)**. The QM network based on 40 mtDNA HV1/HV2 haplotypes

depicted the region-wise clustering patterns of the observed singleton and shared haplotypes **(Figure 1)**. Among the observed haplotypes, we observed h1 to h10 in the Eastern region, of which 50% were unique haplotypes. In the Northern region, we observed h11 to h19 haplotypes, of which 80% were unique haplotypes. The shared haplotype h20 was observed onlyin the South Indian population, and h21 was shared with the South and West Indian population. Likewise, h22 to h26 were uniquely observed (50%) among the West Indians. Moreover, we found some region-wise unique positions in the mtDNA-HV1 viz. variations in 16069, 16169, and 16,206 among the West Indians, 16170, 16181, and 16185 among the North Indians, and 16224 and 16344 among the East Indians **(Figure 2)**. Similarly, we found unique positions in mtDNA-HV2 viz. variation in 41 among the East Indians, 215 and 243 among the West Indians,285 among the North Indians, and 480 among the South Indians.

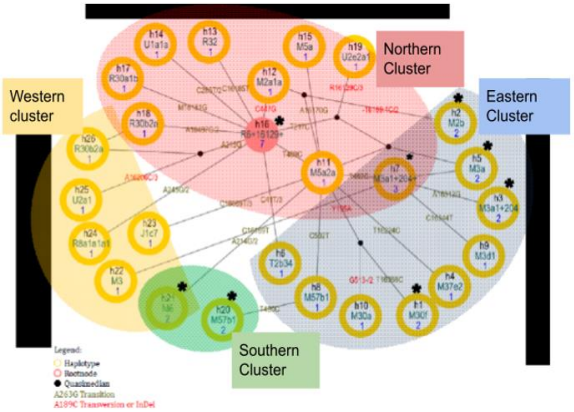
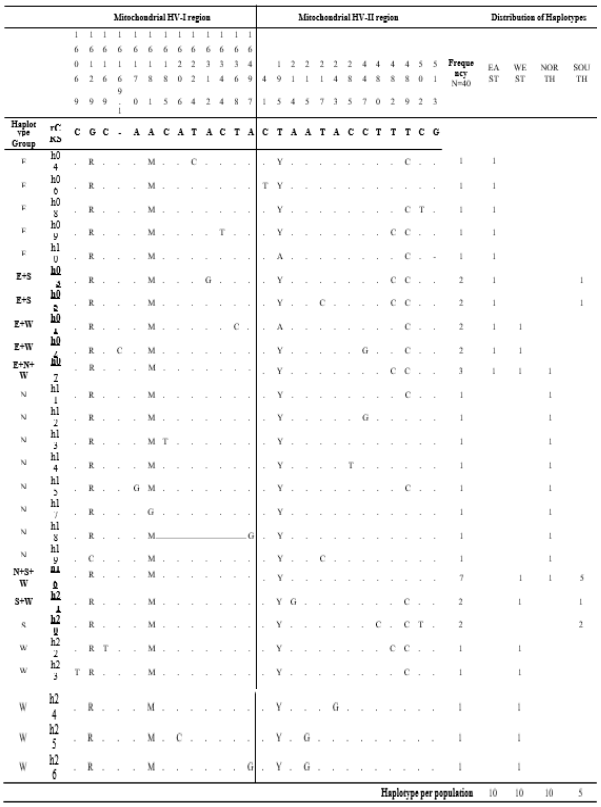


Figure 1: QM network of examined populations based on 40 mtDNA HV1 / HV2 haplotypes. Specific mitochondrial haplogroups are shown in region-specific groupings. *Haplotypes are shared.

Table 4. Distribution of mitochondrial HV-I and HVII haplotypes in our generated 40 individuals



HOTSPOT POSITION	REGION (rCRS position)	EASTERN	NORTHERN	SOUTHERN	WESTERN
16069	HVR-I (16024-16569)				✔
16129			✔	✔	✔
16169					✔
16170			✔		
16181			✔		
16185			✔		
16206					✔
16224		✔			
16312		✔		✔	
16344		✔			
16368		✔			✔
16497			✔		✔
41	HVR-II (1-576)	✔			
195		✔	✔	✔	✔
214				✔	✔
215					✔
217		✔	✔	✔	
243					✔
285			✔		
447		✔	✔		✔
480				✔	
482		✔	✔	✔	✔
489		✔	✔	✔	✔
502		✔		✔	

Color Code: EAST, NORTH, SOUTH, WEST

Figure 2: The distribution of region-specific unique mitochondrial positions in the studied population

Phylogenetic associations of studied Indian populations

We inferred phylogenetic relationships of our studied Indian samples with other Indian populations through the ML method based on mitochondrial HV1 region (nucleotide positions: 16024-16383). We selected the HKY+G substitution model based on the lowest Bayesian Information Criterion scores **(Supplementary Table S5)**. The final tree with the highest log-likelihood (-1313.9585) contained 17 clusters (I-XVII), of which Cluster I was the first South Indian-specific deep cluster. Besides, we exclusively observed East Indian specific clusters (Cluster XV & XIII), North Indian Cluster (Cluster XI), and Central Indian clusters (Cluster XVI&

XVII). Moreover, we observed Eastern & Northern clusters (Cluster X & XII), Eastern & Southern clusters (Cluster II & IX), Northern & Western clusters (Cluster VI), and Southern & Western clusters (Cluster III & V). The remaining four clusters were mixed clusters, of which Northeast Indian sequences were present in three clusters (Cluster IV, VII & VIII), and cluster
XIV contained East, West, North, and Central Indian sequences. However, we did not observeany South Indian samples in these mixed clusters **(Figure 3)**.

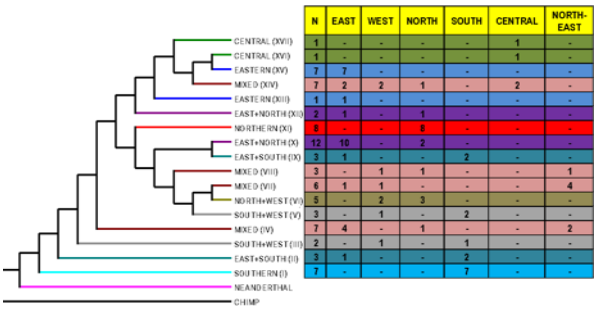


Figure 3: Molecular Phylogenetic analysis by Maximum Likelihood method. Summary information of the evolutionary relationships of 80 mtDNA-HV1 sequences rooted with Neanderthal and Chimpanzee sequences**.** The tree-topology is represented on the left-hand side; N is the number of sequences in each cluster (I - XVII). The table provides the number ofsequences in each cluster taken in this study from the following geographic areas: East, West, North, South, and Northeast India.

Genetic structure evaluation of the studied populations

We evaluated the genetic structure of studied Indian populations by AMOVA. In the Total Population group (model A), we observed 99.97% variance in the "within populations" and 0.03% variance in the "among populations within the group" category. Then these populations were grouped according to geographic criteria (model B), haplogroup pattern (model C), and ML clustering (model D). Among these models, model D showed the maximum variance among groups (2.98) and minimum variance among the populations within groups (-2.51), thus correctly describing the genetic structure of the studied individuals **(Table 5)**.

Table 5. Summary of the AMOVA for estimating the genetic structure of studied Indianpopulations
Assessment of the demographic history of the studied populations

Model	Among groups		Among populations within groups		Within populations	
	Var (%)	P-Value	Var (%)	P-Value	Var (%)	P-Value
[A] Total Group	--	--	0.03	0.026	99.97	--
[B] Geographic Criteria	1.76	0.032	-1.21	0.043	99.45	0.037
[C] Haplogroup pattern	2.25	0.002	-1.91	0.122	99.66	0.034
[D] ML Clustering	**2.98**	< 0.001	**-2.51**	0.031	99.54	0.025

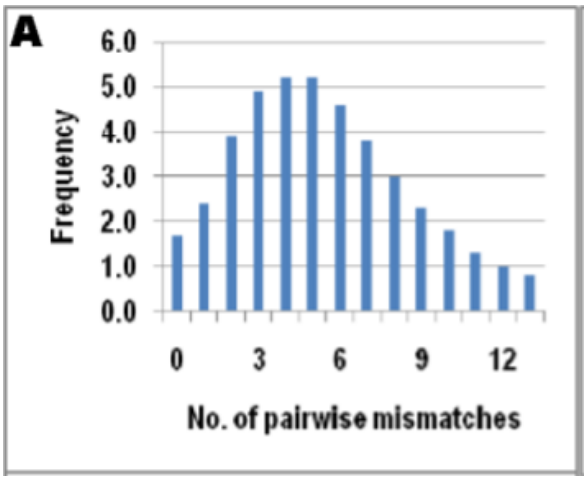
Note. The best model should maximize the variance among groups minimize the variance among population within groups.

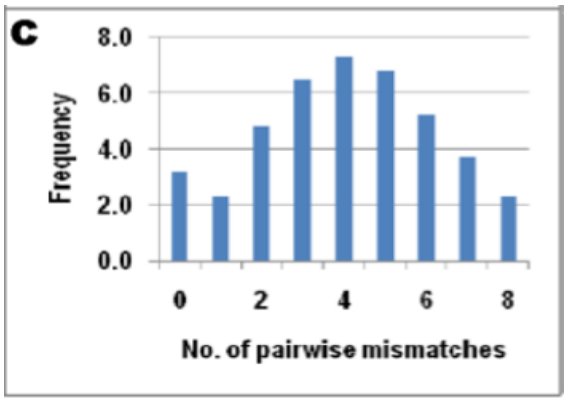
The probable signatures of demographic changes in the studied populations were summarized in **Table 6**. The NORTH group showed higher MPD, Hd, and π than the remaining groups. The unimodal distribution (slightly positively skewed) was observed in all three groups except the SOUTH group, suggesting a prehistoric demographic expansion event in those three groups **(Figure 4)**. A comparatively lower value of r also interpreted a similar observation. Moreover, in these three groups, insignificant Sum of squared deviations (SSD) values, significantly lower R2 statistics values (<0.1), and significant Fs and D statistics were observed, which suggest that they probably underwent spatial expansion. Thus, we observed larger Effective population sizes (Ne>1000) and larger migrant sizes (M>50). We also noticed different expansion times calculated from τ due to probable diverse coalescent histories of these groups, as the NORTH group expanded earlier (~26kya), followed by the EAST group (~22kya) and the WEST group (~15kya). In contrast, we found the lowest Hd, π, and MPD values in the SOUTH group. Also, comparatively higher R2 statistics, higher r-value, slightly multi-modal mismatch distribution pattern, and insignificant Fs and D statistics proposed a relatively stable population size over time in this group. However, we observed a much older demographic expansion time (~28kya), significant SSD value, comparatively smaller Ne (<1000), and M (<50) values.

Table 6. Descriptive statistics of the studied population groups

	Population Groups			
	EAST	WEST	NORTH	SOUTH
Na	10	10	10	10
Nh	9	9	10	7
Nb	21	21	25	11
Hd± s.d.	0.978±0.054	0.978±0.054	1.000±0.045	0.933±0.062
π ± s.d.	0.1121±0.0669	0.1229±0.0727	0.1626±0.0938	0.0813±0.0505
MPD± s.d.	5.267±2.780	5.778±3.020	7.644±3.897	3.822±2.099
SSD	0.0053	0.0317	0.0351	0.0832*
r	0.0183	0.0854	0.1185	0.3047
τ	2.644	1.841	3.148	3.437
θa	3.485	4.760	4.480	0.945
$^{A}Y_{a}$	21,322.58	14,846.77	25,387.10	27,717.74
Ne	1405.24	1919.35	1806.45	381.05
M	534.64	3902.95	8752.56	24.54
Tajima's D	-1.370*	-1.045*	-0.643*	-0.076
Fu's Fs	-3.417*	-3.122*	-4.321*	-1.402
R2	0.0762*	0.0908*	0.0740*	0.1648

Note. Na No. of sequences; Nh No. of haplotypes; Nb No. of polymorphic sites; Hd Haplotype diversity; π ± s.d. Nucleotide diversity ± standard deviation; MPD Mean Pairwise differences ± standard deviation; SSD Sum of squared deviations; r raggedness index; τ tau; θa Initial theta; AYa population expansion age (Years) (A × τ / 2μ); Ne Effective population size (θa/ 2μ); M No. of migrant; R2 Ramos-Onsins and Rozas, R2 statistic; μ polymorphism rate (0.00124 per site per generation); A generation time (20 years); * p-value < 0.05 for R2-statistics, Tajima's D, SSD and p-value < 0.02 for Fu's Fs.





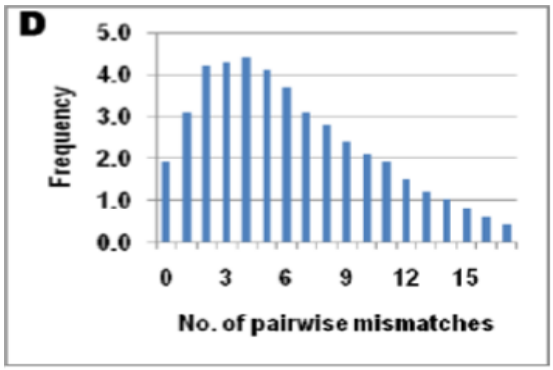


Figure 4: The mismatch distributions of 359bp long mtDNA HV1 sequences of 40 Indian populations. (A) East Indian population, (B) North Indian Population, (C) South Indian population, and (D) West Indian population. The x-axis indicates the number of pairwise nucleotide differences, and the y-axis indicates their simulated frequency

Estimating probable migration events

Finally, we tried to predict the most possible courses of migration in mainland India from the observed mtDNA haplotypes. We evaluated the log-marginal likelihoods to compare different migration models in MIGRATE-n, and the migration model, model#19, showed the maximum (-1101.9) log-marginal likelihood compared to other models **(Supplementary Table S6)**, whose model probability was 56.9% and showed five forward and five backward migration events.
Figure 5 depicts the detailed directional bilateral migration plot for this model.

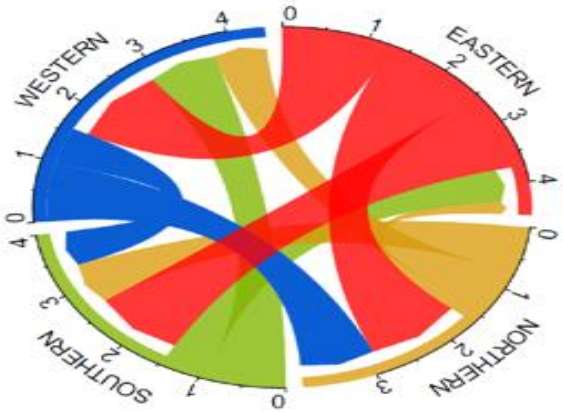


Figure 5: The directional bilateral migration plot for the source-sink migration dynamics among the studied Indian groups based on the mtDNA sequence variations. In this circos plot, the width of migration curves indicates the amount of migration. Both arrows illustrate thedirections of the curves at the end (with the arrowhead) and a height difference. The ticks in theouter circle are showing the proportion of the migration events.

Discussion

In this study, we assessed two Precision ID mtDNA panels on randomly collected unrelated individuals from four zones of India to determine the most suitable panel for challenging samples. We also aimed to evaluate mtDNA HGs diversity and the probable course of migration that led to the current mainland Indian populations in the studied regions. Based on this, we attempted to identify region-specific hotspot mtDNA polymorphisms useful for human identification. The demographic characteristics of the studied individuals showed statistically insignificant differences in the gender-wise distribution of age and region **(Table 2)**, which follows our selection of Age and Ethnicity matched individuals and suggested that we adequatelyperformed random sampling (24).

Detailed descriptive statistics of our sequencing run results showed the effectiveness of both panels for challenging DNA samples. However, our comparative analysis recommended the use of the CR Panel in human identification or any other forensic applications **(Supplementary Table S4)**, as this would be much cheaper and less time-consuming compared to the WG panel for a large number of samples, and more importantly, it can predict mtDNA HGs much quickly and accurately. Additionally, the position-wise distribution of π across the mtDNA D-loop region accurately identified different segments of the control region (Supplementary Figure S3), which,in turn, proves the effectiveness of the studied panel.

Previously, a study proposed that the Eurasian and Oceanian founder mtDNA macro-HG M, N, and R coexisted in South Asia, which were co-migrated intact along the southern coastal route across Arabia to India in one wave after the exit of modern humans from Africa at ~60kya (25), and subsequently differentiated into different sub-HGs in different regions (26). However, the deep-rooted lineages of macro-HG M in existing Indian populations suggested in situ origin of these HGs in South Asia (probably in India), which were not language-specific and dispersed over all the language groups in India (27). A previous autosomal marker-based study also showed the existence of the Ancestral North Indians (ANI) component associated with the Central Asians in the modern Indian populations (28). Similarly, in another study, the presence of the Ancestral Austro-Asiatic (AAA) component was observed in mainland Indian populations (29), and a recent study has indicated that these Austro-Asiatic (AA) speakers have maintained amaternal genetic link between Northeast and Mainland India (17).

Consequently, in this study, the macro-HG M was predominant (78.3%; 18/23) and presented inEast, West, and North India **(Supplementary Table S3)**, which was due to the influence of Indo-European (IE) and AA speakers. Thus, we found that our studied populations clustered as per ML phylogeny, that is, based on linguistic affiliations, accurately predicting the best genetic structure model. Besides, the haplotype analysis showed 69.2% (18/26) unique haplotypes (Table4), which suggested that an elevated mtDNA D-Loop variations exist among these populations (30), and shifting from the endogamous nature towards free-mixing or spreading of IE language within Eurasia (31). Thus, we observed relatively high MPD, π, and Hd as North > West > East **(Table 6)**. The observed mixed ML clusters (Cluster IV, VII, and VIII & XIV) support all these findings **(Figure 3)**. Furthermore, a unimodal mismatch distribution pattern with a slight skewness **(Figure 4)** and significant negative Fu's Fs in these populations suggest a recent expansion from a relatively small population, for which we observed a larger effective population (Ne>1000) and migrant sizes (M>50). Besides, the North Indians showed an older expansion time (~26kya) than the East (~22kya) and West Indians (~15kya), probably due to the forward migrations as NORTH → EAST and SOUTH → EAST, and some backward migrations

as NORTH → WEST and EAST → NORTH, SOUTH & WEST **(Figure 5)**.

Next prevalent macro-HGs R and its descendants (37.5%) were mostly observed in North and West Indian populations, especially the macro-HG U, which was found in North Indians (20%), followed by the West Indian population (10%). It supports that the macro-HG U was brought to India by the founder population which led to the Aryan invasion of India (32). Likewise, we found North Indian Cluster (Cluster XI) and Northern & Western cluster (Cluster VI), containing the IE speakers **(Figure 3)**. Conversely, we uniquely observed ancestral undifferentiated macro-HG R (30%) and W (20%) among the South Indians, the Dravidian speakers, for whichwe noticed a South Indian-specific deep cluster (Cluster I). We also observed macro-HG R in Southern and Western Indian groups, which probably reflected that the undifferentiated macro-HG R migrated to South India with the initial migrants from Africa following the southern coastal route through the Western Indian corridor at ~50kya (33). Consequently, we noticed Southern & Western Indian ML clusters (Cluster III & V) and forward migration as WEST → SOUTH & NORTH, and backward migrations as NORTH & SOUTH → WEST **(Figure 5)**. However, the observed negative value of the variance among the populations within groups in AMOVA may also suggest good admixed populations, lacking an actual subpopulationstructure.

In contrast, none of the haplotypes among South Indians was unique **(Table 4)**, suggesting veryless genetic differentiation in this population bearing the Ancestral South Indians (ASI) component (29), which was congruent with an earlier study on South Indian populations (34). Consequently, we observed the lowest MPD, π, and Hd **(Table 6)**. In addition, other descriptivestatistics for the studied demographic parameters suggest that this group perhaps underwent a sudden demographic expansion in prehistoric times (~28kya), after which they have retained a relatively stable population size over time. We speculate that the prevalence of endogamous nature in this region (35), where free-mixing is infrequent, is probably responsible for such peculiar observations.

Conclusions

In conclusion, this systematic pilot-scale genetic study assessing the Precision ID mtDNA panelson unrelated individuals from four zones of India recommends the Precision ID CR panel for human identification. Additionally, detailed analysis reveals that present mainland Indian populations, especially North, West, and East Indians, are good admixed populations bearing ANI and AAA components. They do not have any real subpopulation structure due to frequent free-mixing and rapid spreading of the Indo-European language across these regions. However, we saw an exception among the South Indians bearing the ASI component, as we observed very less genetic differentiation among them, probably due to their prevailing endogamous societal structure. Ultimately, we have identified 13 region-specific hotspot positions, of which five positions (16069, 16169, 16206, 215 & 243) were West Indian-specific, four positions (16170, 16181, 16185 & 285) were North Indian-specific, three positions (16224, 16344 & 41) were EastIndian-specific and one position (480) was South Indian-specific. However, a further in-depth study with a larger cohort and multiple molecular markers is essential to validate the findings mentioned above and to understand the evolutionary history by evaluating the degree of admixture in current Indian populations.

Acknowledgments

The author gratefully acknowledges the Dean of SBAS, Galgotias University for his motivation.

Conflict of Interest

There is no conflict of interest.

References

[1]. Kundu S, Ghosh SK. Trend of different molecular markers in the last decades for studying human migrations. Gene. 2015; 556(2):81-90. Epub 2014/12/17.

[2]. Klappenbach L. The Process of Evolution. Animals.about.com; 2017 [updated December 19, 201709 December 2021]; Available from: http://animals.about.com/od/evolution/ss/evolution_6.htm.

[3]. Cavalli-Sforza LL, Feldman MW. The application of molecular genetic approaches to the study of human evolution. Nat Genet. 2003; 33:266-75.

[4]. Cuenca D, Battaglia J, Halsing M, Sheehan S. Mitochondrial Sequencing of Missing Persons DNA Casework by Implementing Thermo Fisher's Precision ID mtDNA Whole

Genome Assay. Genes (Basel). 2020; 11(11). Epub 2020/11/08.
[5]. Pakendorf B, Stoneking M. Mitochondrial DNA and human evolution. Annu Rev Genomics Hum Genet. 2005; 6:165-83. Epub 2005/08/30.
[6]. Aquadro CF, Greenberg BD. Human mitochondrial DNA variation and evolution: analysis of nucleotide sequences from seven individuals. Genetics. 1983; 103(2):287-312. Epub 1983/02/01.
[7]. Mondal R, Ghosh SK. Accumulation of mutations over the complete mitochondrial genome in tobacco-related oral cancer from northeast India. Mitochondrial DNA. 2013; 24(4):432-9. Epub 2013/01/29.
[8]. Parson W, Strobl C, Huber G, Zimmermann B, Gomes SM, Souto L, et al. Evaluation of next generation mtGenome sequencing using the Ion Torrent Personal Genome Machine (PGM). Forensic Sci Int Genet. 2013; 7(5):543-9. Epub 2013/08/21.
[9]. Giles RE, Blanc H, Cann HM, Wallace DC. Maternal inheritance of human mitochondrial DNA. Proc Natl Acad Sci U S A. 1980; 77(11):6715-9. Epub 1980/11/01.
[10]. Ghosh SK, Choudhury Y, Mondal R, Laskar RS, Ghosh PR. Individual DNA Barcoding: Promises and Challenges. A Text Book on DNA Barcoding. 1st ed. Kolkata: Books Space; 2012. p. 137-48.
[11]. Faccinetto C, Sabbatini D, Serventi P, Rigato M, Salvoro C, Casamassima G, et al. Internal validation and improvement of mitochondrial genome sequencing using the Precision ID mtDNA Whole Genome Panel. Int J Legal Med. 2021; 135(6):2295-306. Epub 2021/09/08.
[12]. Applied Biosystems. Precision ID mtDNA Panels with the HID Ion S5/HID Ion GeneStudio S5 System; Application Guide. 2021 [06 December 2021]; Available from: https://assets.thermofisher.com/TFS-Assets/LSG/manuals/MAN0017770_PrecisionID_mtDNA_Panels_S 5_UG.pdf.
[13]. Weissensteiner H, Pacher D, Kloss-Brandstätter A, Forer L, Specht G, Bandelt H-J, et al. HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. Nucleic Acids Res. 2016; 44(W1):W58-W63.
[14]. Hasegawa M, Kishino H, Yano T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol. 1985; 22(2):160-74. Epub 1985/01/01.
[15]. Felsenstein J. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. Evolution. 1985; 39(4):783-91.
[16]. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol. 2016; 33(7):1870-4. Epub 2016/03/24.
[17]. Kundu S, Dhar B, Das R, Laskar S, Anil Singh S, Kapfo W, et al. The impact of prehistoric human dispersals on the presence of tobacco-related oral cancer in Northeast India. Gene. 2021; 813:146098. Epub2021/12/25.
[18]. Excoffier L, Laval G, Schneider S. Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evol Bioinform Online. 2007; 1:47-50. Epub 2005/01/01.
[19]. Excoffier L, Smouse PE, Quattro JM. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics. 1992; 131(2):479-91. Epub 1992/06/01.
[20]. Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 2009; 25(11):1451-2.
[21]. Ramos-Onsins SE, Rozas J. Statistical properties of new neutrality tests against population growth. Mol Biol Evol. 2002; 19(12):2092-100. Epub 2002/11/26.
[22]. Beerli P, Palczewski M. Unified framework to evaluate panmixia and migration direction among multiple sampling locations. Genetics. 2010; 185(1):313-26. Epub 2010/02/24.
[23]. Abel GJ. Estimates of Global Bilateral Migration Flows by Gender between 1960 and 2015. Int Migr Rev.2017:1-44. Epub 24 November 2017.
[24]. Ahn EJ, Kim JH, Kim TK, Park JH, Lee DK, Lee S, et al. Assessment of P values for demographic data in randomized controlled trials. Korean J Anesthesiol. 2019; 72(2):130-4. Epub 2018/12/07.
[25]. Mellars P, Gori KC, Carr M, Soares PA, Richards MB. Genetic and archaeological perspectives on the initial modern human colonization of southern Asia. Proc Natl Acad Sci U S A. 2013; 110(26):10699-704. Epub 2013/06/12.
[26]. Kong QP, Sun C, Wang HW, Zhao M, Wang WZ, Zhong L, et al. Large-scale mtDNA screening reveals a surprising matrilineal complexity in east Asia and its implications to the people of the region. Mol Biol Evol. 2011; 28(1):513-22. Epub 2010/08/18.
[27]. Thangaraj K, Chaubey G, Singh VK,

Vanniarajan A, Thanseem I, Reddy AG, et al. In situ origin of deep rooting lineages of mitochondrial Macrohaplogroup 'M' in India. Bmc Genomics. 2006;7.

[28]. Moorjani P, Thangaraj K, Patterson N, Lipson M, Loh PR, Govindaraj P, et al. Genetic evidence for recent population mixture in India. Am J Hum Genet. 2013; 93(3):422-38. Epub 2013/08/13.

[29]. Basu A, Sarkar-Roy N, Majumder PP. Genomic reconstruction of the history of extant populations of India reveals five distinct ancestral components and a complex structure. Proc Natl Acad Sci U S A. 2016; 113(6):1594-9. Epub 2016/01/27.

[30]. Wanjala G, Bagi Z, Kusza S. Meta-Analysis of Mitochondrial DNA Control Region Diversity to Shed Light on Phylogenetic Relationship and Demographic History of African Sheep (Ovis aries) Breeds. Biology (Basel). 2021; 10(8). Epub 2021/08/28.

[31]. Narasimhan VM, Paterson NJ, Moorjani P, Lazaridis I, Mark L, Mallick S, et al. The Genomic Formation of South and Central Asia. bioRxiv. 2018.

[32]. Kivisild T, Bamshad MJ, Kaldma K, Metspalu M, Metspalu E, Reidla M, et al. Deep common ancestry of Indian and western-Eurasian mitochondrial DNA lineages. Current Biology. 1999; 9(22):1331-4.

[33]. Reddy BM, Langstieh BT, Kumar V, Nagaraja T, Reddy ANS, Meka A, et al. Austro-Asiatic Tribes of Northeast India Provide Hitherto Missing Genetic Link between South and Southeast Asia. PLoS ONE. 2007; 2(11):e1141.

[34]. Watkins WS, Thara R, Mowry BJ, Zhang Y, Witherspoon DJ, Tolpinrud W, et al. Genetic variation in South Indian castes: evidence from Y-chromosome, mitochondrial, and autosomal polymorphisms. BMC Genet. 2008; 9:86. Epub 2008/12/17.

[35]. Nakatsuka N, Moorjani P, Rai N, Sarkar B, Tandon A, Patterson N, et al. The promise of discovering population-specific disease-associated genes in South Asia. Nat Genet. 2017; 49(9):1403-7. Epub 2017/07/18.