International Journal of Legal Medicine (2022) 136:465–473
https://doi.org/10.1007/s00414-021-02742-5

ORIGINAL ARTICLE

Forensic characterization of 124 SNPs in the central Indian population using precision ID Identity Panel through next-generation sequencing

Hirak Ranjan Dash[1] · Eduardo Avila[2] · Soumya Ranjan Jena[3] · Kamlesh Kaitholia[1] · Radhika Agarwal[1] · Clarice Sampaio Alho[2] · Ankit Srivastava[4] · Anil Kumar Singh[1]

Received: 22 September 2021 / Accepted: 29 October 2021 / Published online: 8 November 2021

Abstract

With the advent of next-generation sequencing technology, SNP markers are being explored as a useful alternative to conventional capillary electrophoresis–based STR typing. Low mutation rate and short-sized amplicons are added advantages of SNP markers over the STRs. However, to achieve a sufficient level of discrimination among individuals, a higher number of SNPs need to be characterized simultaneously. Hence, the NGS technique is highly useful to analyze a sufficiently higher number of SNPs simultaneously. Though the technique is in its nascent stage, an attempt has been made to assess its usability in the central Indian population by analyzing 124 SNPs (90 autosomal and 34 Y-chromosome) in 95 individuals. Various quality parameters such as locus balance, locus strand balance, heterozygosity balance, and noise level showed a good quality sequence obtained from the Ion GeneStudio S5 instrument. Obtained frequency of SNP alleles ranged from 0.001 to 0.377 in autosomal SNPs. rs9951171 was found to be the most informative SNP in the studied population with the highest PD and lowest MP value. The cumulative MP of 90 SNPs was found to be 4.76698×10^{-37}. Analysis of 34 Y-chromosome SNPs reveals 11 unique haplogroups in 54 male samples with R1a1 as the most frequent haplogroup found in 22.22% of samples. Interpopulation comparison by FST analysis, PCA plot, and STRUCTURE analysis showed genetic stratification of the studied population suggesting the utility of SNP markers present in the Precision ID Identity Panel for forensic demands of the Indian population.

Keywords Next-generation sequencing · SNPs · Central Indian population · Precision ID Identity Panel

Introduction

Present-day forensic DNA analysis relies on the analysis of short tandem repeat (STR) polymorphism using the capillary electrophoresis (CE) technique. This technique is widely accepted for databasing as well as analysis of criminal cases. However, over time, the technology experiences few limitations of its own. The generation of large-sized STR amplicons poses a huge challenge in various biological and compromised samples. Besides, low copy number (LCN) and low template samples, degraded samples, and samples exposed to harsh environmental conditions generate low-quality STR profiles due to the high amplicon size of the conventional STRs. Limited exclusion power in paternity cases and significantly high mutation rate limit the use of STR markers in motherless paternity and in conditions where the relative of the tested man is considered as an alternative father [1].

To overcome the problems associated with STR-based forensic DNA analysis, different other genetic markers such as single-nucleotide polymorphisms (SNPs) and insertion/deletion (indels) have been explored nowadays. SNP markers are the most abundant genetic variations in the human genome occurring once in every 300 nucleotides having a frequency of minor allele > 1% [2]. SNPs in the genome arise due to base substitutions, insertions, or deletions at a single position. Small-sized amplicons, fast genotyping,

✉ Hirak Ranjan Dash
hirakdash@gmail.com

1 DNA Fingerprinting Unit, Forensic Science Laboratory, Bhopal, Madhya Pradesh, India

2 Pontifical Catholic University of Rio Grande Do Sul, Porto Alegre, Brazil

3 Department of Zoology, School of Life Sciences, Ravenshaw University, Cuttack, Odisha, India

4 Institute of Forensic Science and Criminology, Bundelkhand University, Jhansi, UP, India

and low mutation rate of SNP markers provide an added advantage over STR markers in challenging samples. Additionally, recent researches showed other useful applications of SNP markers such as determination of parental lineage, assessment of phenotypic traits, and determination of biogeographical ancestry [3].

Despite many advantages, few limitations are associated with the SNP markers. Due to the biallelic nature of SNPs, they provide less discrimination between individuals in comparison to the multiallelic STR markers. It has been reported that the same level of discrimination is provided by 50–100 autosomal SNPs in comparison to the currently existing core STR loci [4, 5]. As it is a huge task to multiplex such as a huge number of markers for analysis by capillary electrophoresis, next-generation sequencing (NGS) workflow makes it easier. NGS allows simultaneous sequencing of a large number of SNP markers in a huge number of samples at a lesser time. High coverage sequencing also facilitates ultimate single-nucleotide resolution, and the associated automated techniques minimize human intervention. Analysis of SNPs using NGS technology provides added advantage over CE-based technology by generating less artifacts such as stutter products and no noise due to fluorescence. However, lack of sequence interpretation guidelines, high cost, non-availability of population-specific sequence databases, limited study on forensic samples, and complicated workflow limits the current use of NGS technology in forensic DNA analysis. Though many commercial kits are available nowadays for simultaneous analysis of a large number of SNP markers, the Precision ID Identity Panel (Thermo Scientific) allows the simultaneous detection of 90 autosomal SNPs and 34 Y upper-clade SNPs.

The application of NGS-based SNP panels is still in its nascent stage. Various community panels are being developed to predict the biogeographic ancestry, phenotype, and mixture prediction to provide investigative leads [6, 7]. Though the International Society of Forensic Genetics (ISFG) has envisioned the effective use of SNPs to solve complicated forensic cases in conjugation with STRs, a detailed recommendation is still pending in this regard [8]. Thus, the SNP kits available in the market need to be explored in various global populations before commencing their application in real-time forensic case works. In this regard, an attempt has been made to explore the forensic usability of 124 SNP markers present in the Precision ID Identity Panel (Thermo Scientific) in the central Indian population. This is the first report of a South Asian population exploring SNP genomics using Precision ID Identity Panel and Ion GeneStudio™ S5 System (Thermo Scientific, USA). India has a rich migration history and the present-day Indian population is considered to be a mixture of Ancestral North Indians (ANI) related to Central Asians, Middle Easterners, Caucasians, and Europeans and Ancestral South Indians (ASI) mostly the tribal populations [9]. The state of Madhya Pradesh of India is present at the center of the country sharing its boundaries with five other neighboring states. Madhya Pradesh harbors a total of 46 tribal populations, the highest in India. The most common tribal populations inhabiting in Madhya Pradesh include Bhil, Kol, Korku, Sahariya, and Baiga. Thus, the exploration of SNP diversity in the central Indian population can give rise to the representative genetic print of pan India.

Materials and methods

Ethical statement

The samples used in this study were collected from the sample donors after obtaining written informed consent from them. During sample collection and experimentations, the ethical principles mentioned in the Helsinki Declaration were followed strictly. Before commencing this study, approval was obtained from the Ethics Committee at the Maharani Laxmi Bai Medical College, Jhansi, India (Ref. No. 4648/IEC/2020/SC-1).

Sample collection, DNA extraction, and quantification

A total of 95 samples were collected (54 males and 41 females) from the central Indian population. The blood samples were collected from the participants after obtaining written informed consent. The peripheral liquid blood samples collected in K_2EDTA vials were stored at 4 °C until further use. Genomic DNA was extracted from the blood samples by following recommended protocol of the manufacturer using AutoMate Express™ Forensic DNA Extraction System (Thermo Scientific, USA) and PrepFiler Express™ Forensic DNA Extraction Kit (Thermo Scientific, USA). Extracted genomic DNA was subjected to quantification using QuantStudio™ 5 Real-Time PCR System (Thermo Scientific, USA) using Quantifiler® Trio DNA Quantification Kit (Thermo Scientific, USA). Further, the DNA samples were normalized to 1.0 ng/µl in TE buffer and stored at – 20 °C until further use.

Library preparation

DNA libraries were prepared on Ion Chef™ Instrument (Thermo Scientific, USA) using Precision ID Identity Panel (Thermo Scientific, USA) following the recommended protocol. During the library preparation, samples were grouped using Precision ID IonCode™ Barcode Adapters 1–32 Kit (Thermo Scientific, USA). After preparation of libraries, they were quantified using Ion Library TaqMan™

Quantitation Kit (Thermo Scientific, USA) following the manufacturer's guidelines. Libraries were diluted to 30 pM and an equimolar mixture of the libraries was prepared. Further, emulsion PCR and Chip loading were performed on Ion Chef™ Instrument (Thermo Scientific, USA) using the template preparation module.

Sequencing and sequencing parameters

Sequencing was performed on an Ion GeneStudio™ S5 System (Thermo Scientific, USA) using Ion 530™ Chip. For sequencing, a flow rate of 500 was maintained besides other sequencing parameters to be the default. The efficiency of sequencing was assessed by various parameters such as percentage of loading, enrichment, the clonal, final library of Ion Sphere Particle (ISP). Mean read length, percentage of aligned bases, and percentage of mean raw accuracy were also assessed before considering a sequencing result.

Data analysis

The sequencing results were analyzed using Converge™ Software (Thermo Scientific, USA). For allele and genotype calling, the recommended thresholds of various parameters were maintained for minimum allele frequency (0.1), minimum coverage (20), minimum coverage on either strand (10), maximum strand bias (1), MAF(Hom) (95), MAF(Het) (35–65), mean autosomal coverage (10), and Mean Y coverage (20). Further, using Converge module, the random match probability (RMP) of each profile was calculated and the haplogroup of each male profile was determined.

Statistical analysis

Forensic and population parameters such as allele frequencies, observed heterozygosity (Ho), expected heterozygosity (He), match probability (PM), probability of exclusion (PE), polymorphic information content (PIC), power of discrimination (PD), typical paternity index (TPI), linkage disequilibrium (LD), Hardy–Weinberg equilibrium (HWE), and pairwise F_{ST} along with different world populations were performed using STRAF online software (available online at http://cmpg.unibe.ch/shiny/STRAF/) and Arlequin 3.5.2.2. [10, 11]. Population data obtained from 1000 Genomes Project (available online at www.ensembl.org/Home_sapiens/Info/Index) and the Brazilian population [12] were used for the analysis. Multidimensional scaling analysis (MDS) based on pairwise FST distances and principal component analysis (PCA) of worldwide population allele frequencies were calculated using the STRAF online software. Genetic structure among 27 different world populations was carried out using STRUCTURE v.2.3.4 software. Pairwise F_{ST} distances data was used to draw cladogram graphics employing Molecular Evolutionary Genetics Analysis v7.0 (MEGA v7.0) software [13] using the neighbor-joining (N-J) method. Y haplogroup was determined using Converge software and the Y haplogroup frequencies were determined by direct counting.

Results and discussion

Sequencing performance of Precision ID Identity Panel

Three 530 chips were used for sequencing the 124 identity SNP markers. The obtained quality parameters to assess the sequencing performance of samples in three chips are given in Table S1 (Supplementary file). The total number of reads in the sequencing results varied from 10,690,984 to 13,893,772. Loading of ion sphere particles (ISPs) in the chips and percentage enrichment of ISPs were found at a higher value suggesting the efficiency of emulsion PCR and chip-loading process. Similarly, a higher percentage of aligned reads (99.9–100%) and aligned bases (91–94%) with higher mean raw accuracy (99.0–99.1%) suggests the good quality of obtained sequences of the SNP markers. Various statistical parameters were applied to evaluate the efficiency of 124 SNPs using the Precision ID Identity Panel. A two-fold higher average coverage of 90 autosomal SNPs (3606.240 ± 1909.396) was found than that of the 34 upper Y-clad SNPs (1657.883 ± 753.572). A similar result has also been observed when the same panel is sequenced using Ion PGM sequencer in the Brazilian population [12]. The average percentage of + ve strand coverage was found to be in a similar range in both autosomal SNPs (48.721 ± 4.814) and Y-chromosome SNPs (48.498 ± 8.351). Both these values are close to the recommended 50% level of the + ve strand coverage percentage value. This suggests the generation of high-quality sequences determined by the consensus sequences obtained from both positive and negative primers.

Further, the sequencing performance of the Precision ID Identity Panel was assessed marker-wise by calculating the locus balance, locus strand balance, heterozygous balance of autosomal SNPs, and percentage of the noise of all SNPs following Avila et al. [12] (Fig. 1). Locus balance (LB) was calculated by measuring coverage of each locus by average coverage of all loci per sample. An ideal LB value of 1 was achieved by most of the autosomal and Y-chromosome loci. Two autosomal SNPs, i.e., rs1498553 and rs1413212 (0.979), showed the lowest LB values in comparison to all SNPs tested. Similarly, locus strand balance was assessed by estimating forward strand coverage divided by total locus coverage. With a 0.5 optimal value, most of the tested loci showed a similar value with the highest deviation observed at rs876724 (0.596) and rs733164 (0.241). Heterozygous balance (HB) was assessed only for

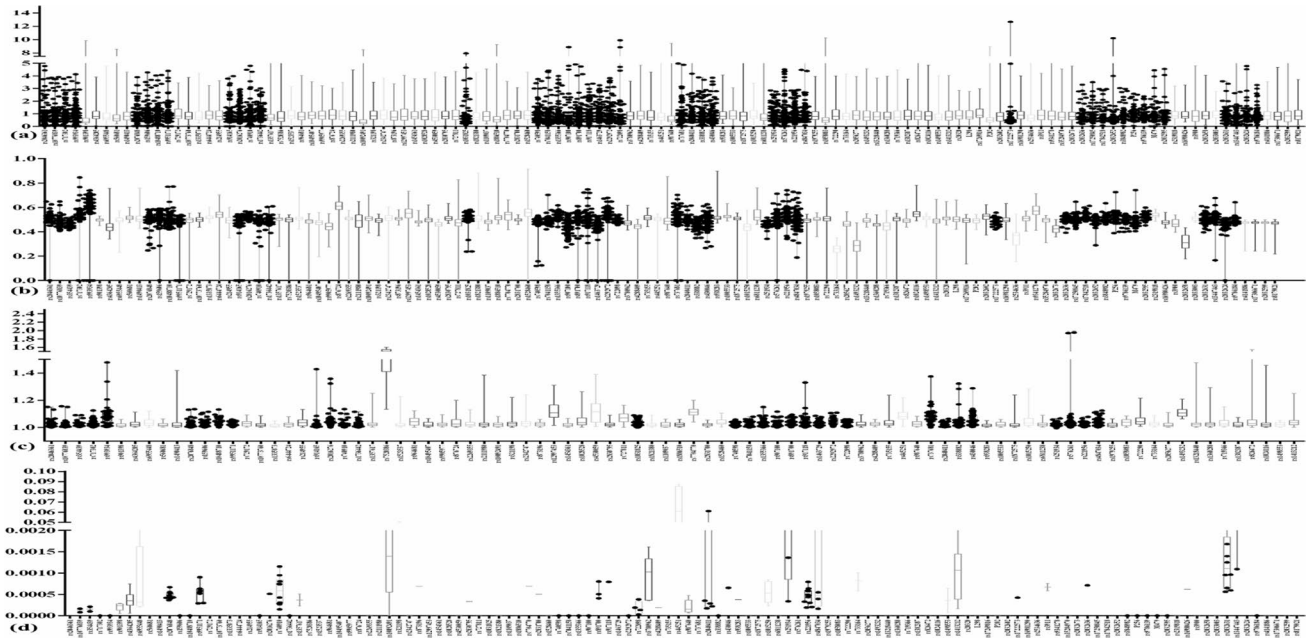

Fig. 1 Marker-wise quality parameters to assess sequencing performance of the Precision ID Identity Panel. Open whiskers represent values for average plus standard deviation, while capped bars indicate maximum or minimum values. **a** Locus balance of all SNPs calculated as coverage of each locus divided by mean coverage of all locus per sample. **b** Locus strand balance of all SNPs calculated as forward strand coverage divided by total locus coverage. **c** Heterozygous balance of autosomal SNPs calculated as coverage ratio of one allele to the other. **d** Percentage of noise of all SNPs calculated as coverage of non-alleles divided by total locus coverage

90 autosomal SNPs included in this study by calculating the coverage ratio of one allele to the other for heterozygous genotypes only. With the optimal value of 1, none of the markers showed significant variation except rs7520386 (1.453). Finally, the noise level of each SNP was evaluated by dividing coverage of non-alleles by total locus coverage, the ideal value of which should be 0. Most of the SNPs showed an absolute 0 noise level after sequencing. However, autosomal SNPs with observed higher noise level include rs445251 (0.062), rs321198 (0.042), rs1109037 (0.032), and rs1031825 (0.028). Similarly, Y-chromosome SNPs with higher noise level includes rs16981290 (0.034), rs2534636 (0.022), and rs2032673 (0.015). However, in comparison to the previous study [12] using the same panel in the Ion PGM™ sequencer, the present study showed significant high-quality data for all the quality parameters tested. This suggests that the use of an automated platform for library preparation, templating, and sequencing by Ion GeneStudio S5 minimizes the noise level and increases the quality of SNP sequences. In a similar line, van der Heijden et al. [14] also advocated the use of Ion Chef™/Ion S5™ workflow for better quality sequencing results and less hands-on time in the laboratory in comparison to the manual library preparation procedure.

Only single source DNA samples were taken into consideration for SNP genotype determination in this study. Mixture analysis was not performed, and it is envisioned that biallelic nature of SNPs will provide less useful information for mixture analysis as they are not as discriminating as STRs and it is immensely difficult to determine the number of contributors in a mixed sample by SNP analysis only. However, as NGS technology provides quantitative data of each allele in terms of coverage, Petrovick et al. [15] used low minor allele frequencies to analyze complex DNA mixtures using NGS. Additionally, another study demonstrated major: minor SNP allele ratio approach to estimate the contributor's proportion in a mixed sample [16]. Hence, the quantitative SNP genotype determination through NGS approach is also helpful in complex mixture analysis.

Forensic parameters of 124 SNPs for the central Indian population

Complete SNP genotypes for 124 SNPs included in this panel are given in Supplementary Table S2. Representative forensic and paternity parameters of the 90 autosomal SNPs are given in Supplementary Table S2. However, the detailed allele frequencies and other forensic parameters of 90 autosomal SNPs are given in Table S3 (Supplementary file). The highest frequency of T (0.831) and lowest frequency of C (0.168) was observed at rs873196 among all the SNPs tested in the studied population. A previous study on the central Indian population reported the allele frequency of STR markers in the range of 0.001 to 0.377

[17]. As most of the SNP markers are biallelic, most of the previous studies have also reported the allele frequencies in a similar line in the US, African-American, and Hispanic populations [18, 19]. Out of 90 polymorphic loci, the highest observed heterozygosity and expected heterozygosity were found at rs964681 (0.610) and rs993934 (0.549), respectively, whereas the lowest observed (0.242) and expected (0.260) heterozygosity were recorded at rs10495407 and rs938283, respectively. In the studied central Indian population, rs938283 showed the lowest PIC, and PD, and the highest PM among 90 SNPs. The highest PIC was observed at rs993934, whereas rs9951171 showed the lowest MP and highest PD. Similarly, rs964681 and rs10495407 showed the highest and lowest PE and TPI among the SNPs tested. *P* values of HWE tests of all loci have been represented in Table S3 (Supplementary File). Out of 90 autosomal SNPs, four of them, i.e., rs1024116, rs1736442, rs1979255, and rs2342747, showed significant deviation ($p < 0.05$) (Table S4). Such HWE deviations of the autosomal SNP markers may be a result of genetic stratification observed in the central Indian population. Statistical significance level was updated to 1.23×10^{-5} after Bonferroni correction, where original significance level (0.05) was divided by the total number of executed pairwise comparisons. Thirty-four SNP marker pairs (highlighted in Table S5) displayed significant linkage disequilibrium. For these 34 SNPs showing linkage disequilibrium, the conservative approach was applied in the statistical analysis (with the use of Bonferroni method and consequent significance level threshold reduction) was adopted to effectively minimize the possibility of such results being observed by chance only. Therefore, statistical support for effective occurrence of linkage disequilibrium in these 34 marker pairs is strong enough to suggest that such phenomenon is actually present in Central India population.

When PM and PE were calculated under the assumption of independence, the cumulative PM and PE of all studied SNP markers in the central Indian population were calculated as 4.76698×10^{-37} and 0.99999692962281, respectively. Previous studies on the central Indian population reported the cPM and cPE values as 1.11×10^{-23} and 0.999998 when 20 CODIS STR loci are used [20]. This signifies higher polymorphism and discriminatory power of the 90 autosomal SNPs, and they can be used in the identification and kinship analysis for forensic purposes. Combinatorial use of STR and SNP markers has been reported to further increase the combined RMP to 3.21×10^{-66} in the Northeastern Peruvian Andes populations [21]. Thus, population-specific analysis of autosomal SNP markers will further explore their usefulness in the human identification process.

Y haplotype analysis using 34 Y upper-clade SNPs

Thirty-four Y upper-clade SNPs were analyzed in 54 male individuals included in this study. Based upon the obtained haplotypes, haplogroups were predicted using Converge software. A total of 11 distinct haplogroups were observed in 54 samples (Fig. 2). Out of which, haplogroup R1a1 was found in 22.22% of samples, whereas haplogroup IJ and R were found in 1.85% of samples. Other predicted haplogroups include H1 (20.37%), R2 (12.96%), O2 & G (9.26% each), J (7.41%), L & C (5.55% each), and Q (3.70%). Haplogroup R1a1 has been widely reported in Eurasia, Central Asia, and the Indian subcontinent based on Y-STR analysis [22]. Other studies have also reported R (38.50%), H (16.10%), L (11.20%), and J (11.10%) haplogroups in Indian populations [23]. Another study on the Indian population also predicted R1a, H, and L haplogroups in 51.5%, 16.2%, and 15.8% samples which was concordant with the predicted haplotypes using 34 Y upper-clade SNPs in the present study [24]. Though the number of samples used in this study to

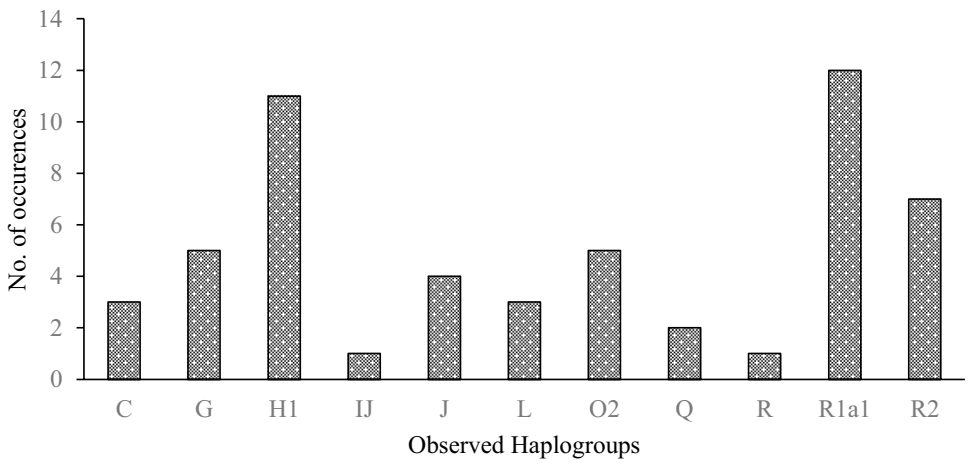

Fig. 2 Predicted Y-chromosome haplogroups in the central Indian population based on 34 upper-clade Y-SNPs ($n = 54$)

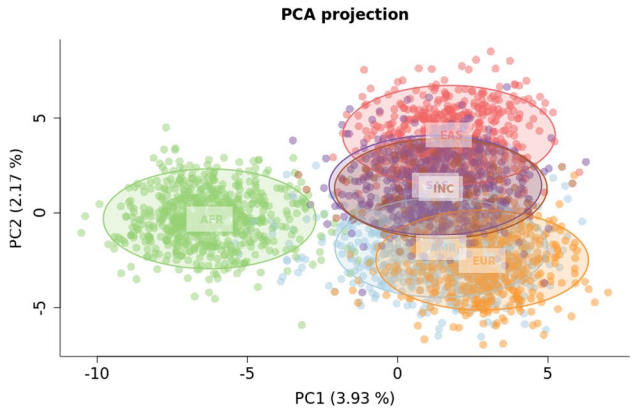

Fig. 3 PCA plot to evaluate the relativeness of the central Indian population with other global populations. The PCA plot was automatically generated by STRAF analysis tool (STRAF Job Id: 1,629,489,656). Each point represents an individual, colored according to the subject's biogeographic origin. For each distinct population, the 95% confidence ellipse is also presented (representing the region containing 95% of samples belonging to the group, drawn from the underlying Gaussian distribution), with the same color patterns as the individuals belonging to the group

predict the haplotypes is less, 34 Y upper clade SNPs are envisioned to predict appropriate haplogroups to understand paternal lineage in the central Indian male individuals.

Interpopulation studies

To study the relatedness of the central Indian population with 27 other global populations, pairwise F_{ST} values were calculated based on 88 overlapped autosomal SNPs (Table S6, Supplementary file) and the obtained heatmap result is given in Fig. S1. Calculated F_{ST} values of the central Indian population (INC) with other global populations ranged from 0.074 to 0.139. Populations such as the Brazilian population (BRA) (0.074073337); the Mexican Ancestry in Los Angeles, CA, USA (MXL) (0.123872582); the African ancestry in SW USA (ASW) (0.126350027); the Toscani in Italia (TSI) (0.126886487); and Colombian in Medellín, Colombia (CLM) (0.127857129) showed close affinity with the central Indian populations. Considering the population data from other regions, the central Indian population showed the most genetic distance with the African population followed by East Asian populations. However, a close similarity was observed with the South Asian population, European populations, and American populations. First and second principal components were evaluated as PCA plot (Fig. 3) and were used to evaluate the relevance of genetic distance among different global populations with respect to the central Indian populations. AFR, AMR, and EUR populations and EAS populations clustered in extreme

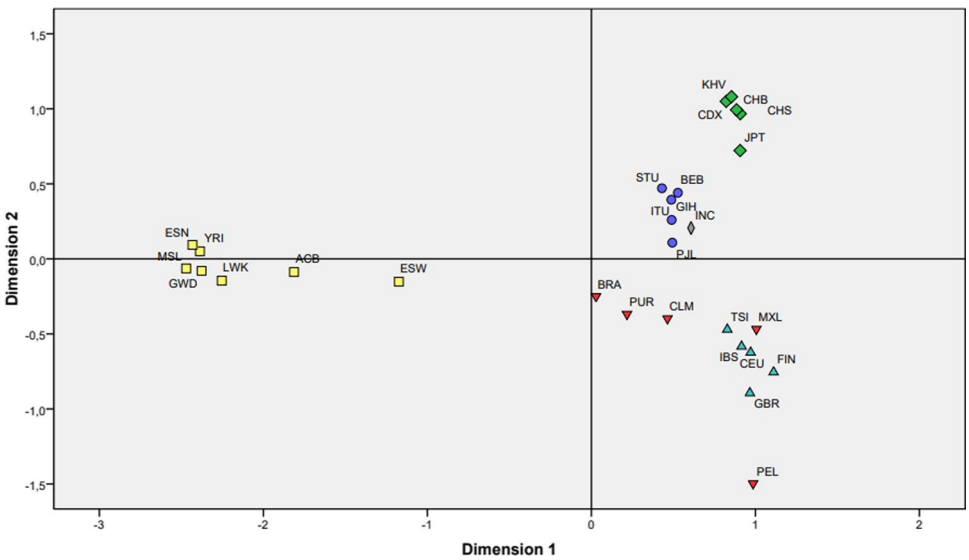

Fig. 4 Genetic distance evaluation for inter population analysis of 27 worldwide populations and the central Indian population, presented as a MDS plot based on pairwise FST values for 88 overlapped autosomal SNPs included in Precision ID Identity Panel (S-Stress = 0.09302 RSQ = 0.96082). Acronyms used are STU, Sri Lankan Tamil in the UK; PJL, Punjabi in Lahore, Pakistan; ITU, Indian Telugu in the UK; GIH, Gujarati Indians in Houston, Texas, USA; BEB, Bengali in Bangladesh; TSI, Toscani in Italia; IBS, Iberian populations in Spain; GBR, British from England and Scotland; FIN, Finnish in Finland; CEU, Utah residents with Northern and Western European ancestry from the CEPH collection; KHV, Kinh in Ho Chi Minh City, Vietnam; JPT, Japanese in Tokyo, Japan; CHS, Han Chinese South, China; CHB, Han Chinese in Beijing, China; CDX, Chinese Dai in Xishuangbanna, China; PUR, Puerto Rican in Puerto Rico; PEL, Peruvian in Lima, Peru; MXL, Mexican Ancestry in Los Angeles CA United States; CLM, Colombian in Medellín, Colombia; YRI, Yoruba in Ibadan, Nigeria; MSL, Mende in Sierra Leone; LWK, Luhya in Webuye, Kenya; GWD, Gambian in Western Division – Mandinka; ESN, Esan in Nigeria; ASW, African ancestry in SW USA; ACB, African Caribbean in Barbados; BRA, the Brazilian population; INC, the Central Indian population

positions, whereas the currently evaluated central Indian population (INC) clustered along with the SAS populations. The PCA plot was automatically generated by STRAF analysis tool (STRAF Job Id: 1,629,489,656). Each point represents an individual, colored according to the subject biogeographic origin. For each distinct population, the 95% confidence ellipse is also presented (representing the region containing 95% of samples belonging to the group, drawn from the underlying Gaussian distribution), with the same color patterns as the individuals belonging to the group. This is following the geographical limitations, genetic, historical, or ethnographic information of the individuals included in this study [25, 26]. Thus, the PCA plot further strengthens the relatedness among the central Indian population with other South Asian populations based on 88 overlapped autosomal SNPs. It shows that the influence of continental origin or ethnicity (as indicated by colors) is not strong enough to provide a good separation of the samples. This information is clear in the graphic, as well as in the low percentage of total variation represented by these components (presented near both axes). Keep in mind that the panel is not designed for ancestry determination, but to human individualization instead. To visualize population-based separation, the MDS image is presented in Fig. 4.

For further characterization of the genetic differences among different global populations with the central Indian population, the Bayesian interference method was studied by STRUCTURE software (Fig. 5). All populations displayed clear distinct clusters as per their geographical origin with the currently studied central Indian population (INC) clustering with other South Asian populations. This further strengthens the utility of autosomal SNPs in population data analysis in the central Indian population. Previous genetic analysis on Indian populations has revealed their descendants from Central Asians, Middle Easterners, Caucasians, and Europeans [9]. Studies have also revealed the African, Chinese, and Great Andamanese ancestry of varied Indian populations based on analysis of 560,123 autosomal SNPs [27]. This is consistent with the history of the Arab slave trade, and Tibeto-Burman language-speaking inhabitants in the Indian population [28]. The presence of Indo-European populations in many Indian states [29] has also supported the European association of the currently studied central Indian population by autosomal SNP analysis.

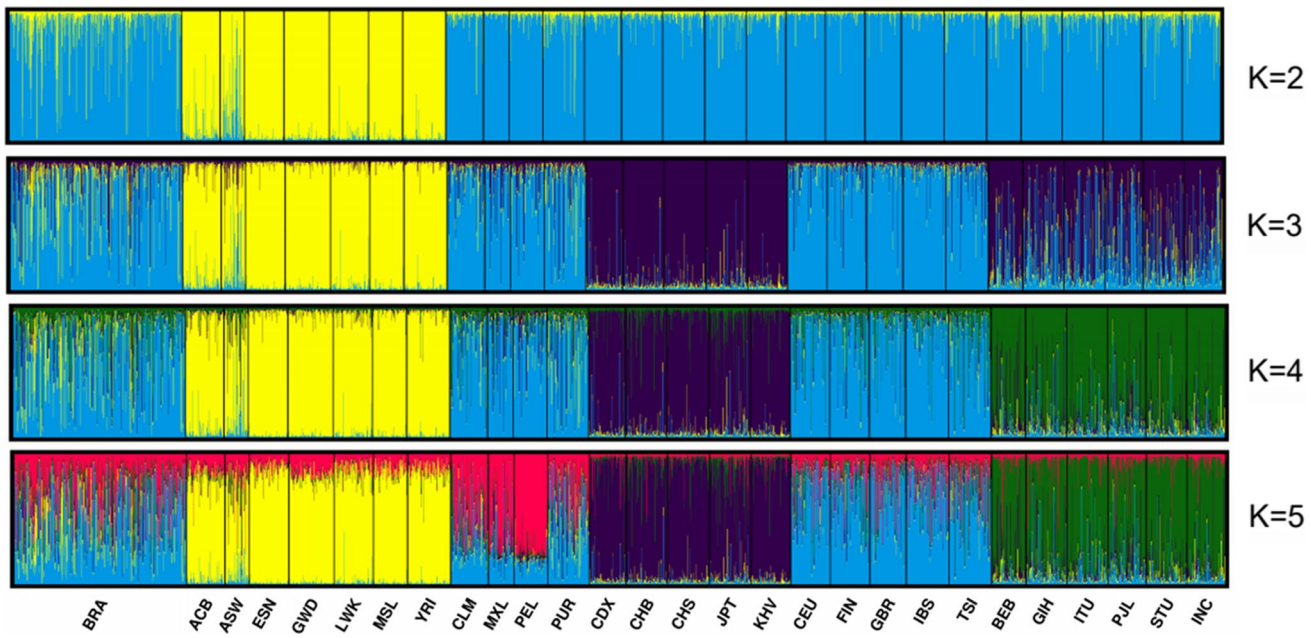

Fig. 5 Population structure of the central Indian population (INC) along with other 27 global populations on the basis of 88 overlapped autosomal SNPs. Acronyms used are STU, Sri Lankan Tamil in the UK; PJL, Punjabi in Lahore, Pakistan; ITU, Indian Telugu in the UK; GIH, Gujarati Indians in Houston, Texas, USA; BEB, Bengali in Bangladesh; TSI, Toscani in Italia; IBS, Iberian populations in Spain; GBR, British from England and Scotland; FIN, Finnish in Finland; CEU, Utah residents with Northern and Western European ancestry from the CEPH collection; KHV, Kinh in Ho Chi Minh City, Vietnam; JPT, Japanese in Tokyo, Japan; CHS, Han Chinese South, China; CHB, Han Chinese in Beijing, China; CDX, Chinese Dai in Xishuangbanna, China; PUR, Puerto Rican in Puerto Rico; PEL, Peruvian in Lima, Peru; MX,: Mexican Ancestry in Los Angeles, CA, USA; CLM, Colombian in Medellín, Colombia; YRI, Yoruba in Ibadan, Nigeria; MSL, Mende in Sierra Leone; LWK, Luhya in Webuye, Kenya; GWD, Gambian in Western Division – Mandinka; ESN, Esan in Nigeria; ASW, African ancestry in SW USA; ACB, African Caribbean in Barbados; BRA, the Brazilian population; INC, the Central Indian population

Conclusions

Forensic DNA analysis is witnessing a paradigm shift from the conventional capillary electrophoresis–based STR typing to next-generation sequencing of SNP markers. Low mutation rate and short-sized amplicons are added advantages of SNP markers over the STRs. However, to achieve a sufficient level of discrimination among individuals, a higher number of SNPs need to be characterized simultaneously. Hence, the NGS technique is highly useful to analyze a sufficiently higher number of SNPs simultaneously. Though the technique is in its nascent stage, an attempt has been made to assess its usability in the central Indian population by analyzing 124 SNPs (90 autosomal and 34 Y-chromosome) in 95 individuals. Various quality parameters such as locus balance, locus strand balance, heterozygosity balance, and noise level showed a good-quality sequence obtained from the Ion GeneStudio S5 instrument. Obtained frequency of SNP alleles ranged from 0.001 to 0.377 in autosomal SNPs. rs9951171 was found to be the most suitable SNP in the studied population with the highest PD and lowest MP value. The cumulative MP of 90 SNPs was found to be 4.76698×10^{-37}. Analysis of 34 Y-chromosome SNPs reveals 11 unique haplogroups in 54 male samples with R1a1 as the most frequent haplogroup found in 22.22% of samples. Interpopulation comparison by FST analysis, PCA plot, and STRUCTURE analysis showed genetic stratification of the studied population suggesting the utility of SNP markers present in the Precision ID Identity Panel for forensic demands of the Indian population.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00414-021-02742-5.

Acknowledgements The authors are highly acknowledged to Director, State Forensic Science Laboratory, Sagar, M. P., India, and Joint Director, Regional Forensic Science Laboratory, Bhopal, M. P., India for providing infrastructure to carry out the research work. Our sincere thanks to Dr. Atima Agrawal, Dr. Neeraj Chauhan, Dr. Sanjib Dey, and the entire technical team of Thermo Scientific for their constant technical support during the research work.

Declarations

Ethical The samples used in this study were collected from the sample donors after obtaining written informed consent from them. During sample collection and experimentations, the ethical principles mentioned in the Helsinki Declaration were followed strictly. Before commencing this study, approval was obtained from the Ethics Committee at the Maharani Laxmi Bai Medical College, Jhansi, India (Ref. No. 4648/IEC/2020/SC-1).

Conflict of interest The authors declare no competing interests.

References

1. Schneider PM (2012) Beyond STRs: the role of diallelic markers in forensic genetics. Transfus Med Hemother 39:176–180. https://doi.org/10.1159/000339139
2. Nelson MR, Marnellos G, Kammerer S, Hoyal CR, Shi MM, Cantor CR, Braun A (2004) Large-scale validation of single nucleotide polymorphisms in gene regions. Genome Res 14:1664–1668. https://doi.org/10.1101/gr.2421604
3. Yousefi S, Abbassi-Daloii T, Kraaijenbrink T, Vermaat M, Mei H, van't Hof P, van Iterson M, Zhernakova DV, Claringbould A, Franke L, 't Hart LM, Slieker RC, van der Heijden A, de Knijff P, BIOS consortium, 't Hoen PAC (2018) A SNP panel for identification of DNA and RNA specimens. BMC Genomics 19:90. https://doi.org/10.1186/s12864-018-4482-7
4. Budowle B, van Daal A (2018) Forensically relevant SNP classes. BIOTECHNIQUES 44:5. https://doi.org/10.2144/000112806
5. Chakraborty R, Stivers DN, Su B, Zhong Y, Budowle B (1999) The utility of short tandem repeat loci beyond human identification: implications for development of new DNA typing systems. Electrophoresis 20:1682–1696
6. Diepenbroek M, Bayer B, Schwender K, Schiller R, Lim J, Lagacé R, Anslinger K (2020) Evaluation of the Ion AmpliSeq™ PhenoTrivium panel: MPS-based assay for ancestry and phenotype predictions challenged by casework samples. Genes 11:1398. https://doi.org/10.3390/genes11121398
7. Oldoni F, Bader D, Fantinato C, Wootton SC, Lagacé R, Kidd KK, Podini D (2020) A sequence-based 74plex microhaplotype assay for analysis of forensic DNA mixtures. Forensic Sci Int Genetics 49:102367. https://doi.org/10.1016/j.fsigen.2020.102367
8. Butler JM, Budowle B, Gill P, Kidd KK, Phillips C, Schneider PM, Vallone PM, Morling N (2008) Report on ISFG SNP Panel Discussion. Forensic Sci Int Genet Suppl Ser 1:471–472. https://doi.org/10.1016/j.fsigss.2007.10.159
9. Moorjani P, Thangaraj K, Patterson N, Lipson M, Loh PR, Govindaraj P, Berger B, Reich D, Singh L (2013) Genetic evidence for recent population mixture in India. Am J Hum Genet 93:422–438. https://doi.org/10.1016/j.ajhg.2013.07.006
10. Excoffier L, Lischer HE (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour 10:564–567. https://doi.org/10.1111/j.1755-0998.2010.02847.x
11. Gouy A, Zieger M (2017) STRAF - a convenient online tool for STR data evaluation in forensic genetics. Forensic Sci Int Genet 30:148–151. https://doi.org/10.1016/j.fsigen.2017.07.007
12. Avila E, Felkl AB, Graebin P, Nunes CP, Alho CS (2019) Forensic characterization of Brazilian regional populations through massive parallel sequencing of 124 SNPs included in HID ion Ampliseq Identity Panel. Forensic Sci Int Genet 40:74–84. https://doi.org/10.1016/j.fsigen.2019.02.012
13. Kumar S, Stecher G, Tamura K (2016) MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol 7:1870–1874. https://doi.org/10.1093/molbev/msw054
14. van der Heijden S, de Oliveira SJ, Kampmann ML, Børsting C, Morling N (2017) Comparison of manual and automated AmpliSeq™ workflows in the typing of a Somali population with the Precision ID Identity Panel. Forensic Sci Int Genet 31:118–125. https://doi.org/10.1016/j.fsigen.2017.09.009
15. Petrovick MS, Boettcher T, Fremont-Smith P, Peragallo C, Ricke DO, Watkins J, Schwoebel E (2020) Analysis of complex DNA mixtures using massively parallel sequencing of SNPs with low minor allele frequencies. Forensic Sci Int Genet 46:102234. https://doi.org/10.1016/j.fsigen.2020.102234
16. Ricke DO, Fremont-Smith P, Watkins J, Stankiewicz S, Boettcher T, Schwoebel E (2019) Estimating Individual Contributions to

Complex DNA SNP Mixtures. J Forensic Sci 64:1468–1474. https://doi.org/10.1111/1556-4029.14030
17. Dixit S, Shrivastava P, Kumawat RK, Kaitholia K, Dash HR, Sharma H, Choubey G (2019) Forensic genetic analysis of population of Madhya Pradesh with PowerPlex Fusion 6C^{TM} Multiplex System. Int J Leg Med 133:803–805. https://doi.org/10.1007/s00414-019-02017-0
18. Vallone PM, Decker AE, Butler JM (2005) Allele frequencies for 70 autosomal SNP loci with U.S. Caucasian, African-American, and Hispanic samples. Forensic Sci Int 149:279–286. https://doi.org/10.1016/j.forsciint.2004.07.014
19. Kiesler KM, Vallone PM (2013) Allele frequencies for 40 autosomal SNP loci typed for US population samples using electrospray ionization mass spectrometry. Croat Med J 54:225–231. https://doi.org/10.3325/cmj.2013.54.225
20. Dash HR, Rawat N, Vajpayee K, Shrivastava P, Das S (2021) Useful autosomal STR marker sets for forensic and paternity applications in the Central Indian population. Ann Hum Biol 48:37–48. https://doi.org/10.1080/03014460.2021
21. Guevara EK, Palo JU, King JL, Buś MM, Guillén S, Budowle B, Sajantila A (2021) Autosomal STR and SNP characterization of populations from the Northeastern Peruvian Andes with the ForenSeq^{TM} DNA Signature Prep Kit. Forensic Sci Int Genet 52:102487. https://doi.org/10.1016/j.fsigen.2021.102487
22. Sharma S, Rai E, Sharma P, Jena M, Singh S, Darvishi K, Bhat AK, Bhanwer AJS, Tiwari PK, Bamezai RNK (2009) The Indian origin of paternal haplogroup R1a1* substantiates the autochthonous origin of Brahmins and the caste system. J Hum Genet 54:47–55. https://doi.org/10.1038/jhg.2008.2
23. Mahal DG, Matsoukas IG (2018) The geographic origins of ethnic groups in the Indian subcontinent: exploring ancient footprints with Y-DNA haplogroups. Front Genet 9:4. https://doi.org/10.3389/fgene.2018.00004
24. Singh M, Sarkar A, Nandineni MR (2018) A comprehensive portrait of Y-STR diversity of Indian populations and comparison with 129 worldwide populations. Sci Rep 8:15421. https://doi.org/10.1038/s41598-018-33714-2
25. Tätte K, Pagani L, Pathak AK, Kõks S, Duy BH, Ho XD, Sultana GNN, Sharif MI, Asaduzzaman M, Behar DM, Hadid Y, Villems R, Chaubey G, Kivisild T, Metspalu M (2019) The genetic legacy of continental scale admixture in Indian Austroasiatic speakers. Sci Rep 9:3818. https://doi.org/10.1038/s41598-019-40399-8
26. Debortoli G, Abbatangelo C, Ceballos F, Fortes-Lima C, Norton HL, Ozarkar S, Parra EJ (2020) Jonnalagadda M (2020) Novel insights on demographic history of tribal and caste groups from West Maharashtra (India) using genome-wide data. Sci Rep 10:10075. https://doi.org/10.1038/s41598-020-66953-3
27. Reich D, Thangaraj K, Patterson N, Price AL, Singh L (2009) Reconstructing Indian Population History. Nat 461:489–494. https://doi.org/10.1038/nature08365
28. Thangaraj K, Ramana GV, Singh L (1999) Y-chromosome and mitochondrial DNA polymorphisms in Indian populations. Electrophoresis 20:1743–1747
29. Debortoli G, Abbatangelo C, Ceballos F, Fortes-Lima C, Norton HL, Ozarkar S, Parra EJ, Jonnalagadda M (2020) Novel insights on demographic history of tribal and caste groups from West Maharashtra (India) using genome-wide data. Sci Rep 10:10075. https://doi.org/10.1038/s41598-020-66953-3