STR genetic diversity from the Human Genome Diversity Project (HGDP) populations.

Tamara Soledad Frontanilla[a]*, Guilherme Valle-Silva[b], Jesús Ayala[c], Celso Teixeira Mendes-Junior[a].

[a]Departamento de Genética, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, 14049-900, Ribeirão Preto, SP, Brazil.

[b]Departamento de Química, Laboratório de Pesquisas Forenses e Genômicas, Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo, 14040-901, Ribeirão Preto, SP, Brazil.

[c]Universidad de la Integración de las Americas. Asunción, Paraguay.

*Corresponding author:
Tamara Soledad Frontanilla
tfronta@gmail.com

ORCID ID
Tamara Soledad Frontanilla: 0000-0002-6873-7813
Guilherme Valle-Silva: 0000-0002-0062-9162
Jesus Ayala: 0000-0002-7065-6879
Celso Teixeira Mendes-Junior: 0000-0002-7337-1203

ABSTRACT

The Human Genome Diversity Project (HGDP) studied 54 worldwide populations comprising seven population groups: African, American, Central South Asian, East Asian, European, Middle Eastern, and Oceanian. This study aimed to perform a comprehensive genotyping analysis of STRs commonly used in forensic and population genetic studies from the Human Genome Diversity Project dataset and to publish it as an open-access STR database to contribute to future forensic genetics studies. A set of 22 STR markers were analyzed using high-coverage Whole-Genome Sequencing data from BAM files available at the International Genome Sample Resource. HipSTR was used to call genotypes from 929 samples from all 54 population samples. To validate our results, we directly compared our NGS-based and CE-based genotypes on 16 STRs available in Rosenberg lab (Stanford University) dataset. Also, the allele frequencies estimated were compared with the data stored at the SPSmart STR browser. Forensic parameters, allele frequencies, and Hardy-Weinberg equilibrium adherence were calculated for each population. Principal Coordinate Analysis (PCoA), the Analysis of Molecular Variance (AMOVA), and clustering analysis were used to evaluate population structure. The D21S11 marker could not be detected in the present study. The average successful calling rate was 90.27%, ranging from 58.56% (Penta D) to 97.85% (D3S1358). Comparing both databases, the average number of identical genotypes was 97.44%. In conclusion, this investigation offers a population genetics perspective based on a comprehensive genotyping analysis of STR commonly used in the forensic genetics field, concerning the whole Human Genome Diversity Project dataset. Except for Penta D and Penta E, all genotypes and allele frequencies presented in this study are supported by (a) previous reports that certify HipSTR's reliability, (b) the comparison between CE-derived and NGS-derived genotypes, (c) frequency data reports from worldwide populations, including the large pop.STR database, and (d) the conclusions achieved by our population genetics analysis that corroborates current knowledge regarding modern human demographic history.

Keywords: HipSTR; allele frequencies; forensic genetics; worldwide population; bioinformatics.

INTRODUCTION

The Human Genome Diversity Project (Almarri et al., 2020; Bergström et al., 2020; Cavalli-Sforza, 2005) (HGDP) is a collaboration of scientists worldwide to create a database of different world populations. It was started in 1990 by Stanford University's Morrison Institute (Cann et al., 2022; Rosenberg, 2006). The project initially had some ethical issues concerning indigenous populations (Dodson et al., 1999), who are considered vulnerable and might be exploited (Cavalli-sforza, 2005). In 1994, after a few years of discussion, the US National Research Council (NRC) of the National Academy of Sciences (NAS) recommended that the HGDP should proceed because of the countless scientific benefits, but always carrying out the necessary care and consent. This project studied 54 worldwide populations comprising seven population groups: African, American, Central South Asian, East Asian, European, Middle Eastern, and Oceanian (Bergström et al., 2020).

There are many international collaborative genome-wide studies, such as The Human Genome Project (HGP) (Birney, 2021) the Haplotype Map (HapMap) project (1000 Genomes Project Consortium et al., 2015), the Human Genome Diversity Project (Cavalli-Sforza, 2005), and the 1000 Genomes Project (1000 Genomes Project Consortium et al., 2015). The first two focus on mapping and sequencing genes to discover their relationship with different diseases. However, the HGDP and the 1000 Genomes Project are more interested in understanding the extent of genetic variation between humans. Although The 1000 Genomes Project has produced an extensive catalog of human genetic variation, the HGDP contains samples of underrepresented human populations or isolated indigenous populations that are necessary to better understand the demographic history and introgression of Neanderthals and Denisovans’ DNA into modern human genomes (Callaway, 2019; Degioanni et al., 2019; Demeter et al., 2022).

Next generation sequencing (NGS) (Behjati et al., 2013), also known as massively parallel sequencing or deep sequencing, is a revolutionary technology that allows the sequencing of millions of small DNA fragments in parallel. Specifically developed bioinformatics tools are used to piece together these fragments using the human reference genome as a backbone. NGS can evaluate thousands or even millions of *loci* simultaneously compared with just a few dozen *loci* detected by PCR and electrophoresis (Bonneville *et al.*, 2020).

Genome-wide studies, including exome and/or whole-genome sequencing, are becoming more and more common worldwide for diagnosing rare genetic diseases and predicting possible forthcoming conditions. Such datasets allow for the analysis of more complex genetic regions that are usually left aside, such as Short Tandem Repeats (STR) markers. STRs are composed of consecutive repetitive units of 2-6 base pairs that form series with lengths of up to 100 nucleotides or even more (Fan; 2007). Typically, capillary electrophoresis (CE) is the technique used to genotype these markers after PCR amplification. However, recent articles (Ganschow et al., 2018; Gymrek et al., 2012; Valle-Silva et al., 2022; Warshauer et al., 2013; Willems et al., 2017) showed that specific bioinformatic tools could successfully genotype these markers from NGS data.

Haplotype inference and phasing for STRs (Willems et al., 2017; Gordon et al., 2017) (HipSTR) is a bioinformatic tool developed for calling STR markers specifically from Whole Genome Sequencing (WGS). It was created to process hundreds of samples at once, making it suitable to deal with large databases. Moreover, HipSTR learns locus-specific PCR stutter models using an EM algorithm, employing a specialized hidden Markov model to align reads to candidate alleles while accounting for STR artifacts and using phased SNP haplotypes to genotype and phase STR. HipSTR showed high accuracy in previous studies, demonstrating a 98.8% consistency compared with capillary electrophoresis in 118 samples (Halman; Oshlack, 2020). Valle-Silva et al. (2022) compared three software to genotype STR markers from NGS data showing more than 97% calling accuracy between them.

This study aimed to perform a comprehensive genotyping analysis of STRs commonly used in population genetics studies from the Human Genome Diversity Project dataset and to publish it as an open-access STR database.

METHODOLOGY

Genotype Calling

The population sample consisted of 929 individuals from the Human Genome Diversity Project (HGDP) panel, distributed across 54 worldwide populations that compose seven population groups: Africa (*n*=104), Americas (*n*=61), Central South Asia (*n*=197), East Asia (*n*=223), Europe (*n*=155), Middle East (*n*=161) and Oceania (*n*=28). These populations are described by Bergstrom et al. (Bergström et al., 2020) (Table 1). The CRAM files containing sequence data from these 929 samples are available at the International Genome Sample Resource, divided into two datasets: one presented by Mallick et al. (Mallick et al., 2016) (https://www.internationalgenome.org/data-portal/data-collection/hgdp), and the other by Bergström et al. (2020) (https://www.internationalgenome.org/data-portal/data-collection/sgdp).

Table 1. Population samples from the Human Genome Diversity Project (HGDP) used in this study (*n*=929).

Population	Subpopulation	Nomenclature	Number of individuals
Africa	BantuKenya	AFR001	11
	BantuSouthAfrica	AFR002	8
	Biaka	AFR003	22
	Mandenka	AFR004	22
	Mbuti	AFR005	13
	San	AFR006	6
	Yoruba	AFR007	22
America (Amerindians)	Colombian	AMR008	7
	Karitia	AMR009	12
	Maya	AMR010	21
	Pima	AMR011	13
	Surui	AMR012	8
Central/South Asia	Balochi	CSA013	24
	Brahui	CSA014	25
	Burusho	CSA015	24
	Hazara	CSA016	19
	Kalash	CSA017	22
	Makrani	CSA018	25
	Pathan	CSA019	24
	Sindhi	CSA020	24

	Uygur	CSA021	10
East Asia	0xi	EAS022	8
	Cambodian	EAS023	9
	Dai	EAS024	9
	Daur	EAS025	9
	Han	EAS026	33
	Hezhen	EAS027	9
	Japanese	EAS028	27
	Lahu	EAS029	8
	Miao	EAS030	10
	Mongolian	EAS031	9
	NorthernHan	EAS032	10
	Oroqen	EAS033	9
	She	EAS034	10
	Tu	EAS035	10
	Tujia	EAS036	9
	Xibo	EAS037	9
	Yakut	EAS038	25
	Yi	EAS039	10
Europe	Adygei	EUR040	16
	Basque	EUR041	23
	BergamoItalian	EUR042	12
	French	EUR043	28
	Orcadian	EUR044	15
	Russian	EUR045	25
	Sardinian	EUR046	28
	Tuscan	EUR047	8
Middle East	Bedouin	MES048	46
	Druze	MES049	42
	Mozabite	MES050	27
	Palestinian	MES051	46
Oceania	Bougainville	OCE052	11
	PapuanHighlands	OCE053	9
	PapuanSepik	OCE054	8

All samples were sequenced in high-coverage as described by Mallick et al. (MALLICK; LI; LIPSON; MATHIESON *et al.*, 2016) and Bergström et al. (2020). This coverage depth provides a reliable opportunity to genotype STR markers accurately despite their large sizes (i.e., repetitive sequences encompassing up to 130 bp).

A total of 22 *loci* commonly referenced in forensic practice were analyzed: CSF1PO, D1S1656, D2S441, D2S1338, D3S1358, D5S818, D7S820, D8S1179, D10S1248, D12S391, D13S317, D16S539, D18S51, D19S433, D21S11, D22S1045, FGA, Penta D, Penta E, TH01, TPOX, and vWA. The HipSTR (Willems et al., 2017) algorithm was run for each individual to genotype the 22 STRs based on the human reference genome GRCh38, and using the BED file hg38.hipstr_reference.bed with the flanking regions available in the HipSTR repository (https://github.com/HipSTR-Tool/HipSTR-references/). A minimum of 8 reads were accepted to obtain reliable genotypes, and the 15% stutter model as a calling filter was used.

The genotypes for each marker were calculated using three parameters provided in the output VCF file: the reference allele of each marker, the period (i.e., the length of each STR repeat unit), and the base pair differences (GB) in comparison with the reference allele (Willems et al., 2017). Nomenclature adjustments were made for D19S433, Penta D, Penta E, and vWA following Valle-Silva et al. (Valle-Silva et al., 2022) recommendations to couple with the nomenclature established by the International Society for Forensic Genetics (ISFG) (Gettings; et al., 2019). By using the IGV software (Robinson et al., 2017; Thorvaldsdóttir et al., 2013) and the HipSTR VizAln function (Willems et al., 2017), we have previously demonstrated that the alleles provided by HipSTR for these four markers (D19S433, Penta D, Penta E, and vWA) needed nomenclature adjustment to avoid a shift of some base pairs in allele calling (Valle-Silva et al., 2022). The adjustments consisted of removing two repeat units from all D19S433 and vWA alleles called by HipSTR, including one repeat unit into all Penta D alleles, and removing two nucleotides from all Penta E alleles (Gettings et al., 2019).

Statistical Analysis

Forensic parameters [Match Probability (MP), Power of Discrimination (PD), Power of Exclusion (PE), and Polymorphism Information Content (PIC)], allele frequencies, and Hardy-Weinberg equilibrium adherence were estimated for each population group using GenAlEx 6.5 (Peakall, 2012) and STRAF 2.5.1 software (Gouy; Zieger, 2017).

To explore the distribution of genetic diversity across populations of different ethnic backgrounds, the Principal Coordinate Analysis (PCoA), the Analysis of Molecular Variance (AMOVA), and clustering analysis were done using GenAlEx 6.5 (Peakall, 2012), Arlequin 3.5 (Excoffier; Lischer, 2010), and STRUCTURE 2.3.4 (Hubisz et al., 2009) software, respectively. The STRUCTURE analysis was performed for *k* ranging from 3 to 7, applying the correlated allele frequency model and 200.000 burn-in steps followed by 200.000 Markov Chain Monte Carlo interactions in 10 independent runs. The results from the runs with the largest "Estimated Ln Probability of Data" [LnP(D)] were selected and are depicted in bar plots created with Clumpak (Kopelman et al., 2015).

Genotype validation

We used two validation methodologies to verify the reliability of genotype data generated by HipSTR. The first one consisted of a direct comparison with CE-derived genotypes available in the Rosenberg's lab (Stanford University) dataset available at: https://rosenberglab.stanford.edu/data/algeehewittEtAl2016/HGDPmicrosatsIncludingCODIS.stru. The dataset is composed of the data published by Algee-Hewitt *et al.* (2016) (Algee-Hewitt et al., 2016) and Rosenberg et al. (2005) (Rosenberg et al., 2005). We used 865 individuals and 16 STR markers present in both the NGS and CE datasets for this validation.

In a secondary validation attempt, the allele frequencies estimated in the present study were compared with allele frequency data from the same seven major population groups (African, European, Middle Eastern, Central South Asian, East Asian, Oceanian, and American) stored at the SPSmart STR browser (Amigo et al., 2009; Fernandez et al., 2009) (Pop.STR). For this comparison, pairwise F_{ST} was estimated using the Arlequin software (Excoffier; Lischer, 2010).

RESULTS

STR genotypes established for each individual from the HGDP dataset using HipSTR are available in Supplementary Table 1 as an open-access database. The D21S11 marker was excluded because we failed in genotyping it. The mean coverage for genotype calling ranged from 29.765 (Penta D) to 53.869 (D3S1358) (Table 2).

Table 3 shows the allele frequencies and forensic parameters for the whole HGDP dataset, while Supplementary Table 2 presents these same parameters for each of the seven major population groups studied. The average successful calling rate was 90.27%, ranging from 58.56% (Penta D) to 97.85% (D3S1358) (Table 3). HipSTR failed in genotyping the Penta D alleles smaller than five repeats. Moreover, Penta E was in H-W disequilibrium in half (27) of the 54 population samples (Table 4). Thus, these markers were excluded from all interpopulation statistical analyses performed in the present study (Analysis of Molecular Variance, STRUCTURE analysis, and PCoA). It is noteworthy that the D22S1045 was monomorphic in a small (*n*=13) Amerindian population sample of Mexico (Pima); however, this is due to a lack of genetic diversity in this locus rather than genotyping errors.

Table 2. Average coverages obtained for each STR using the HipSTR tool.

Maker	Lowest value	Median	Highest value	Mean	Standard deviation
CSF1PO	22	47	158	47.704	14.659
D1S1656	21	49	138	49.757	15.574
D2S441	19	48	125	48.643	14.656
D2S1338	28	52	115	47.994	22.810
D3S1358	23	53	134	53.869	15.041
D5S818	12	44	117	44.856	13.887
D7S820	18	41	118	41.309	12.603
D8S1179	24	50	137	50.962	15.589
D10S1248	20	43	105	43.073	14.102
D12S391	23	53	122	48.841	22.612
D13S317	15	40	120	40.670	13.406
D16S539	25	48	123	47.867	14.269
D18S51	29	52	154	50.016	22.480
D19S433	22	47	104	45.073	16.933
D22S1045	8	46	121	39.821	22.946
FGA	28	56	132	50.909	23.812
PentaD	16	38	130	29.765	27.295
PentaE	11	39	119	30.427	23.215
TH01	17	40	115	40.778	12.295
TPOX	19	37	101	35.364	14.620
vWA	23	44	173	43.086	22.804

Table 3. Allele frequencies and the forensic parameters estimated for each marker in the whole HGDP dataset.

Allele	CSF1PO	D1S1656	D2S441	D2S1338	D3S1358	D5S818	D7S820	D8S1179	D10S1248	D12S391	D13S317	D16S539	D18S51	D19S433	D22S1045	FGA	PentaD	PentaE	TH01	TPOX	vWA
5																	0.002	0.100			
5.2														0.001							
6	0.001						0.001										0.007		0.204	0.012	
7	0.010					0.019	0.018		0.001		0.001						0.014	0.108	0.237	0.006	
8	0.012	0.003	0.001			0.014	0.167	0.006	0.002		0.150	0.013					0.040	0.053	0.141	0.457	
9	0.024	0.002	0.003			0.055	0.086	0.005	0.001		0.099	0.172		0.001			0.252	0.036	0.265	0.135	
9.1			0.005				0.001														
9.3			0.001																0.133		
10	0.261	0.006	0.215			0.132	0.256	0.110	0.001		0.087	0.116	0.005	0.014	0.014		0.165	0.085	0.018	0.076	
10.1			0.001																		
10.3									0.001												
11	0.275	0.071	0.351			0.314	0.277	0.063	0.008		0.269	0.291	0.014	0.012	0.179		0.194	0.232	0.002	0.276	0.001
11.1							0.001														
11.2													0.001	0.001							
11.3			0.056																		
11.4																		0.001			
12	0.356	0.085	0.087		0.001	0.288	0.162	0.117	0.052		0.284	0.262	0.087	0.072	0.018		0.132	0.205	0.001	0.038	
12.1												0.001									
12.2													0.001	0.008							
12.3			0.009											0.001							
13	0.054	0.095	0.032		0.003	0.163	0.030	0.235	0.258		0.084	0.125	0.128	0.259	0.005		0.134	0.088			0.001
13.1													0.001								
13.2													0.002	0.044							
13.3		0.001	0.002																		
14	0.006	0.117	0.211		0.057	0.014	0.003	0.238	0.296	0.001	0.026	0.018	0.181	0.286	0.046		0.040	0.057			0.131
14.2														0.060							
14.3		0.002																			
15	0.001	0.183	0.023		0.351	0.001		0.166	0.223	0.022	0.001	0.002	0.157	0.093	0.326		0.009	0.029			0.087
15.2														0.089							

15.3	0.021														
16	0.185	0.003	0.027	0.284	0.001	0.048	0.124	0.023	0.121	0.035	0.251	0.001	0.007	0.006	0.230
16.2				0.001					0.001	0.023					
16.3	0.043														
17	0.064	0.001	0.138	0.201		0.011	0.033	0.092	0.126	0.001	0.154	0.001	0.002		0.268
17.1								0.001							
17.2										0.001					
17.3	0.080							0.008							
18	0.011		0.081	0.094		0.002	0.002	0.210	0.089		0.007	0.012			0.188
18.2										0.001		0.001			
18.3	0.023							0.016							
19			0.163	0.009				0.205	0.051		0.001	0.053			0.078
19.1								0.001							
19.2								0.001				0.002			
19.3	0.008							0.006							
20			0.136					0.141	0.021			0.078			0.014
20.2												0.001			
21			0.036					0.099	0.010			0.122			0.001
21.2												0.003			
22			0.053					0.081	0.004			0.219			
22.2												0.005			
22.3								0.001							
23			0.226					0.056	0.001			0.158			
23.2												0.004			
24			0.086					0.027	0.001			0.178			
24.2								0.001				0.003			
24.3												0.001			
25			0.045					0.007				0.102			
25.2												0.004			
26			0.009					0.001				0.041			
26.2								0.001							
27												0.008			

28																0.004					
28.1																0.001					
30																0.001					

	CSF1PO	**D1S1656**	**D2S441**	**D2S1338**	**D3S1358**	**D5S818**	**D7S820**	**D8S1179**	**D10S1248**	**D12S391**	**D13S317**	**D16S539**	**D18S51**	**D19S433**	**D22S1045**	**FGA**	**PentaD**	**PentaE**	**TH01**	**TPOX**	**vWA**
N	905	903	908	793	909	907	907	903	886	806	899	902	823	852	778	800	544	627	905	850	803
Na	10	18	16	11	9	10	11	11	13	22	9	9	20	19	10	24	13	12	8	7	10
Ho	0.728	0.843	0.675	0.755	0.724	0.731	0.763	0.792	0.721	0.814	0.750	0.763	0.854	0.764	0.666	0.754	0.778	0.440	0.706	0.664	0.762
He	0.726	0.883	0.773	0.864	0.744	0.770	0.795	0.829	0.777	0.864	0.800	0.787	0.877	0.822	0.773	0.860	0.832	0.859	0.794	0.689	0.808
MP	0.126	0.025	0.081	0.037	0.109	0.085	0.071	0.051	0.081	0.033	0.068	0.076	0.028	0.053	0.087	0.036	0.050	0.058	0.070	0.151	0.062
PE	0.473	0.681	0.391	0.519	0.466	0.478	0.532	0.584	0.462	0.625	0.509	0.532	0.703	0.534	0.377	0.516	0.558	0.140	0.438	0.374	0.531
PD	0.874	0.975	0.919	0.963	0.891	0.915	0.929	0.949	0.919	0.967	0.932	0.924	0.972	0.947	0.913	0.964	0.950	0.942	0.930	0.849	0.938
PIC	0.677	0.873	0.741	0.850	0.702	0.735	0.765	0.807	0.742	0.850	0.772	0.755	0.864	0.801	0.738	0.844	0.811	0.845	0.762	0.642	0.781
CMP	**3.72.E-26**																				
CPE	0.999999676																				

N: number of samples; Na: number of alleles; Ho: Observed Heterozygosity; He: Expected Heterozygosity; MP: match probability; PE: power of exclusion; PD: power of discrimination; PIC: polymorphism information content; CMP: combined match probability; CPE combined power of exclusion.

Table 4. Probabilities of adherence to Hardy-Weinberg equilibrium proportions for each STR in all 54 subpopulations analyzed in the HGDP. Significant *p*-values (α = 0.05) are in boldface.

Pop	CSF1PO	D1S1656	D2S441	D2S1338	D3S1358	D5S818	D7S820	D8S1179	D10S1248	D12S391	D13S317	D16S539	D18S51	D19S433	D22S1045	FGA	PentaD	PentaE	TH01	TPOX	vWA
AFR001	0.539	0.248	0.766	0.628	0.673	0.433	0.763	0.988	0.290	0.174	0.837	0.300	0.521	0.682	0.340	0.223	0.423	0.336	0.191	0.505	0.842
AFR002	0.712	0.350	0.824	0.229	0.824	0.930	0.440	0.621	0.374	0.888	0.273	0.261	0.438	0.161	**0.017**	0.177	0.157	0.116	0.673	0.704	0.390
AFR003	0.074	0.812	**0.008**	0.636	0.181	0.537	0.529	0.900	0.947	0.379	0.808	0.929	0.562	0.363	**0.042**	0.787	0.450	0.059	0.553	0.834	0.985
AFR004	0.742	0.461	0.662	0.201	0.916	0.967	0.576	0.630	0.964	0.441	0.724	0.648	0.141	0.992	0.245	0.864	0.526	**0.004**	0.702	0.136	0.603
AFR005	0.207	0.256	0.085	**0.012**	0.617	0.938	0.519	0.427	0.669	0.751	0.631	0.518	0.289	0.353	0.256	0.121	0.823	**0.038**	0.569	0.607	0.900
AFR006	0.556	0.548	0.227	0.654	0.868	0.678	0.556	0.868	0.166	0.874	0.054	0.226	0.174	0.626	0.767	0.062	0.766	0.207	0.995	0.502	0.393
AFR007	0.291	0.317	0.981	0.305	0.059	**0.011**	0.381	0.289	0.605	0.385	0.425	0.257	0.812	0.631	0.708	0.283	0.595	**0.007**	0.682	0.149	0.561
AMR008	0.914	0.678	0.914	0.694	0.466	0.950	0.312	0.556	**0.021**	0.735	0.479	0.776	0.511	0.575	**0.034**	0.282	0.387	0.116	0.626	0.152	0.626
AMR009	0.122	0.724	0.197	0.308	0.785	0.942	0.950	0.711	0.615	0.535	**0.044**	0.568	**0.031**	0.763	0.451	0.820	0.841	0.083	0.376	0.792	**0.043**
AMR010	0.999	0.446	0.990	0.687	0.524	0.254	0.707	0.403	0.461	**0.001**	0.283	0.795	0.149	0.210	**0.011**	0.138	0.678	**0.008**	0.795	0.463	0.566
AMR011	0.800	0.645	0.199	0.461	0.637	**0.001**	0.229	0.853	0.337	0.949	0.307	0.615	0.460	0.200	-	0.546	0.711	**0.003**	0.623	**0.028**	0.827
AMR012	0.820	0.498	0.686	0.978	0.983	**0.028**	0.719	0.836	0.726	0.557	0.947	0.217	0.545	0.117	0.054	0.628	0.542	**0.046**	0.733	0.409	0.442
CSA013	0.532	0.273	0.165	0.350	**0.000**	0.215	0.869	0.718	0.478	0.306	0.915	0.224	0.737	0.063	0.565	**0.011**	0.892	**0.000**	0.655	0.861	0.802
CSA014	0.909	0.407	0.073	0.334	**0.010**	0.809	0.973	0.688	0.920	0.870	0.588	0.735	0.620	0.106	0.557	0.329	0.363	0.614	0.325	0.726	0.845
CSA015	0.615	0.611	0.617	0.746	**0.043**	**0.037**	0.695	0.144	0.875	0.135	0.922	0.374	0.999	0.091	0.765	**0.000**	0.584	**0.011**	0.366	0.797	0.342
CSA016	0.652	0.940	**0.009**	0.521	0.849	0.669	0.180	0.917	**0.006**	0.457	0.290	0.073	0.764	0.867	0.163	**0.039**	0.272	0.136	0.194	0.423	0.611
CSA017	**0.873**	0.966	**0.079**	0.799	0.467	0.985	0.486	0.983	0.608	0.840	0.632	0.361	0.577	0.064	0.001	**0.088**	0.607	0.022	0.882	0.823	0.810
CSA018	0.759	0.565	0.997	0.920	0.442	0.847	0.876	0.645	0.320	**0.030**	0.154	0.249	0.733	0.192	0.221	0.596	0.248	**0.001**	0.912	0.949	0.343
CSA019	0.984	0.787	0.908	0.488	0.593	0.716	0.085	0.857	0.465	**0.007**	0.350	0.144	0.510	0.716	0.845	0.345	0.559	0.153	0.764	0.834	0.908
CSA020	0.976	0.585	0.797	0.792	0.436	0.707	0.124	0.689	0.939	0.930	0.795	0.168	0.191	0.107	0.937	0.485	0.081	**0.001**	0.955	0.461	0.684
CSA021	0.093	0.803	0.884	0.585	0.379	0.777	0.264	0.899	0.606	0.609	0.756	0.258	0.214	0.678	0.678	0.084	0.509	**0.013**	**0.029**	0.540	0.399
EAS022	0.460	0.611	0.421	0.262	0.587	0.440	0.760	0.269	0.896	0.718	0.154	0.741	0.453	0.466	0.570	0.505	0.396	**0.019**	0.311	0.572	0.212
EAS023	0.779	0.917	0.173	0.732	0.231	0.543	0.676	0.493	0.656	0.779	0.511	0.523	0.469	0.902	0.516	0.245	0.744	0.109	0.103	0.895	0.521
EAS024	0.213	0.607	0.182	**0.006**	0.948	0.134	0.544	0.101	0.947	0.866	0.685	0.467	0.233	0.172	0.320	0.899	0.849	**0.003**	0.837	0.896	0.744
EAS025	0.083	0.621	0.423	0.656	0.878	0.197	**0.016**	**0.009**	0.620	0.326	0.137	0.276	0.482	0.312	0.320	0.388	0.338	**0.006**	0.586	0.322	0.815
EAS026	0.930	0.823	0.831	0.868	0.960	0.693	0.569	0.727	0.121	0.239	0.920	**0.025**	0.989	0.961	0.610	0.540	0.562	**0.000**	0.587	0.867	0.829
EAS027	0.677	0.103	0.969	0.373	0.652	0.794	0.479	0.577	0.183	0.413	0.206	0.524	0.932	0.661	0.720	0.704	0.804	0.229	0.099	0.726	0.518
EAS028	0.712	**0.008**	0.215	0.815	0.481	0.792	0.243	0.257	0.690	0.988	0.674	0.649	0.301	0.404	0.850	0.928	0.373	**0.000**	0.223	0.623	0.622

EAS029	0.631	0.572	0.423	0.651	0.947	0.777	0.850	0.820	0.478	0.725	0.389	0.147	0.739	0.353	0.796	0.649	0.641	0.132	0.824	0.685	0.226
EAS030	0.374	0.839	0.345	0.354	**0.019**	0.567	0.714	0.452	0.317	0.432	0.695	0.832	0.247	0.922	0.538	0.469	0.540	0.256	0.329	0.606	0.738
EAS031	0.376	0.967	0.799	0.641	0.638	0.878	0.922	0.223	0.895	0.529	0.734	0.575	0.168	0.676	0.572	0.375	0.891	0.062	0.941	0.789	0.761
EAS032	0.202	0.581	0.571	0.192	0.398	0.704	0.288	0.198	0.717	0.418	0.548	0.595	0.154	0.332	0.815	0.956	0.870	0.103	0.497	0.111	0.402
EAS033	0.791	0.389	0.268	0.626	0.599	0.387	0.895	0.789	0.615	0.602	0.639	0.877	0.252	0.120	0.581	0.459	0.936	**0.022**	0.806	0.946	0.465
EAS034	0.546	0.287	0.363	0.440	0.506	0.481	0.375	0.787	0.974	0.966	0.800	0.925	0.582	0.538	0.274	0.258	0.507	0.177	0.374	0.120	0.570
EAS035	0.681	0.413	0.590	0.637	0.379	0.427	0.537	0.924	0.611	0.497	0.453	**0.050**	0.816	0.431	0.799	0.459	0.711	0.227	0.501	0.587	0.534
EAS036	0.119	0.171	0.369	0.402	0.656	0.685	0.544	0.956	0.855	0.804	0.402	0.922	0.854	0.232	0.366	0.451	0.677	0.125	0.723	0.472	0.370
EAS037	0.265	0.276	0.342	0.412	0.532	0.509	0.187	0.077	0.812	0.442	0.752	0.382	0.279	0.768	0.544	0.078	0.716	0.282	0.552	0.716	0.891
EAS038	0.977	**0.005**	0.801	0.400	0.928	0.902	0.924	0.354	0.951	0.942	0.871	0.592	0.966	0.964	0.238	0.791	0.646	0.215	0.768	0.849	0.866
EAS039	0.769	0.354	0.564	**0.003**	0.393	0.617	0.798	0.617	0.581	0.147	0.883	0.273	0.879	0.163	0.154	0.615	0.134	0.062	0.064	0.856	0.780
EUR040	0.641	0.462	0.385	0.918	0.369	0.487	0.564	0.235	0.804	0.796	0.983	0.953	0.168	0.917	**0.016**	0.488	0.658	**0.002**	0.430	0.084	0.793
EUR041	0.440	0.326	0.469	0.743	0.563	**0.033**	0.727	0.722	0.841	0.705	0.253	0.800	0.472	0.810	0.988	0.140	0.593	0.118	0.606	0.941	0.735
EUR042	0.091	0.073	0.292	0.376	0.434	0.470	**0.005**	0.218	0.851	0.525	**0.026**	0.512	0.689	0.540	0.907	0.221	0.220	**0.032**	0.216	0.625	0.245
EUR043	0.730	0.709	0.153	0.341	0.305	0.970	0.100	0.809	0.920	**0.049**	0.139	0.653	0.833	**0.001**	0.089	0.839	0.214	0.053	0.936	0.655	0.671
EUR044	**0.005**	0.396	0.607	0.413	0.093	0.258	**0.045**	0.772	0.833	0.666	0.080	0.966	0.170	0.950	0.966	0.721	0.402	**0.038**	0.345	0.425	0.086
EUR045	0.774	0.312	0.732	**0.018**	0.889	0.432	0.747	0.304	0.335	0.551	0.209	0.087	0.689	0.504	**0.042**	0.693	0.841	**0.000**	0.503	0.998	0.241
EUR046	0.515	0.137	0.065	0.728	0.530	0.393	0.665	0.223	0.414	0.104	0.822	0.961	0.421	0.692	0.680	0.331	0.172	**0.002**	0.619	0.333	0.234
EUR047	0.792	0.496	0.307	0.227	0.977	0.340	0.834	0.326	0.878	0.569	0.735	0.502	0.550	0.104	0.436	0.177	0.371	0.086	0.848	0.949	0.851
MES048	0.145	0.890	0.061	0.068	0.907	**0.008**	0.364	0.840	0.511	**0.021**	0.440	0.635	0.256	0.238	0.753	**0.009**	0.944	**0.000**	0.781	0.519	0.641
MES049	0.144	**0.000**	1.000	0.270	0.230	0.864	0.689	**0.000**	0.484	**0.000**	0.618	0.976	0.226	0.374	**0.006**	0.232	0.377	**0.000**	0.115	0.085	0.282
MES050	0.342	0.566	0.241	0.851	0.667	0.230	0.478	0.123	0.865	0.462	0.990	0.648	0.081	0.948	0.714	0.499	0.556	**0.021**	0.728	**0.050**	0.685
MES051	0.973	**0.024**	0.276	0.628	0.950	0.921	**0.005**	0.357	0.954	0.987	0.746	0.076	0.570	0.760	0.658	**0.000**	0.599	**0.023**	0.814	0.974	0.833
OCE052	0.636	0.867	0.181	0.432	0.463	0.857	0.979	0.214	0.472	0.780	0.420	0.263	0.594	0.566	0.635	0.164	0.565	0.085	0.678	0.377	0.175
OCE053	0.254	0.451	0.719	0.111	0.799	0.671	0.864	0.633	0.936	0.338	0.552	0.517	0.964	0.764	0.244	0.361	0.558	**0.000**	0.381	0.557	0.453
OCE054	0.849	0.154	0.686	0.674	0.974	0.062	0.272	0.276	0.647	0.160	0.183	0.412	0.371	0.916	0.108	0.164	0.729	0.157	0.757	0.677	0.867

The Principal Coordinates Analysis (PCoA) shows a good differentiation between major biogeographic populations at both continental (Figure 1) and subcontinental (Figure 2) scales. For Figure 1, all subpopulations were grouped, revealing four different population clusters. As expected, the African, Amerindian, and Oceanian populations were placed separately (in different quadrants), while the European and Asian populations were clustered together, revealing a similar genetic composition. The two principal coordinates account for 70.42% of the variance. In Figure 2, although the two first coordinates account for only 24.14% of the variance, the distribution of the 54 subpopulations was consistent with what was observed in Figure 1, resulting in four different and well-defined clusters. However, in the cluster with the European and Asian populations, one may observe an overlapping of populations from the four groups, mainly European, Middle Eastern, and Central South Asian populations, corroborating their shared ancestry and similar genetic compositions.

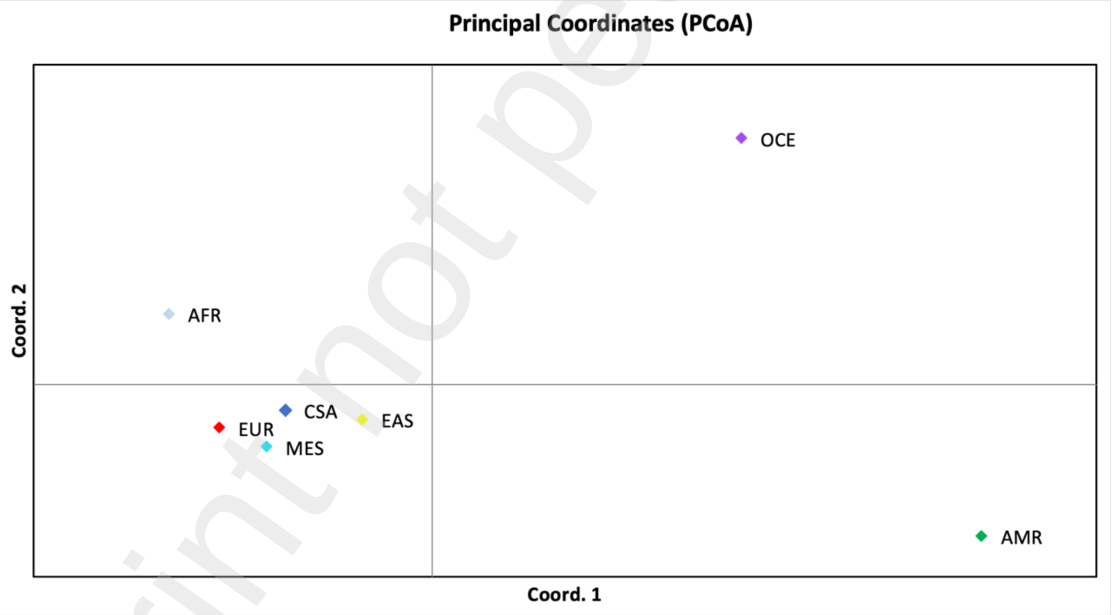

Figure 1. Principal Coordinates Analysis (PCoA) based on autosomal STR data from the 7 major populations of the HGDP. Coordinates 1 and 2 account for 40.66% and 29.76% of the variance, respectively. Penta D and Penta E markers were excluded from this analysis. (AFR: African; CSA: Central South Asia; EAS: East Asia; EUR: European; MES: Middle East; OCE: Oceania).

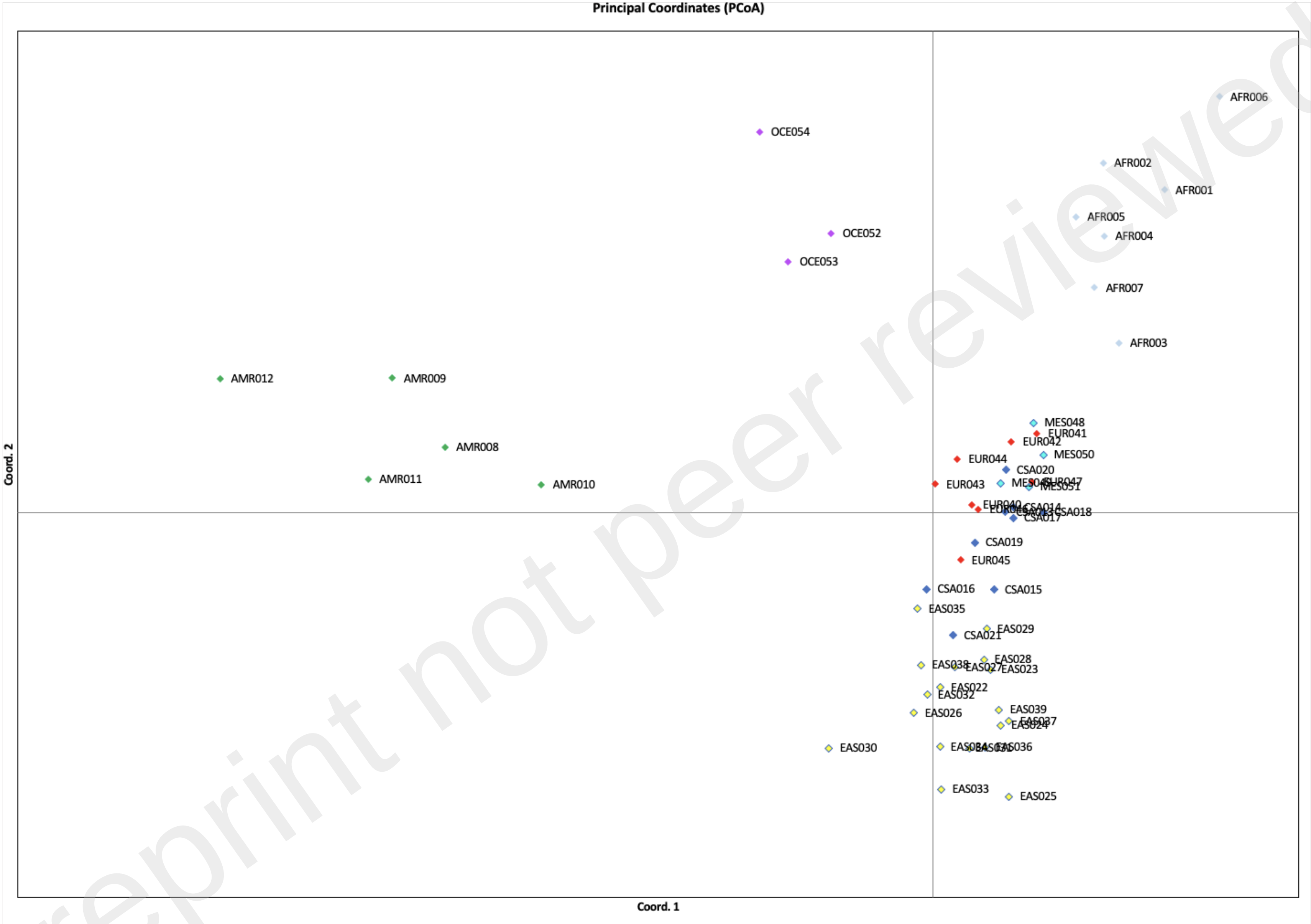

Figure 2. Principal Coordinates Analysis (PCoA) based on autosomal STR data from the 54 sub-populations of the HDGP. Coordinates 1 and 2 account for 13.48% and 10.66% of the variance, respectively. Penta D and Penta E markers were excluded from this analysis. (AFR: African; CSA: Central South Asia; EAS: East Asia; EUR: European; MES: Middle East; OCE: Oceania).

Similar results were obtained with the STRUCTURE analysis. Figure 3 depicts STRUCTURE results from runs obtained with *k* ranging from 3 to 7. With *k* = 5, one may observe that African, Amerindian and Oceanian groups mainly present their own clusters, while European and Asian populations display their shared ancestry, especially European, Middle Eastern and Central South Asian populations. By analyzing *k* = 7, it is possible to observe that the Central South Asian populations are highly heterogeneous with each other but also present evident differences when compared to European and Middle Eastern populations. Although minor differences arise with *k* = 7, European and Middle Eastern populations are very similar in all *k*.

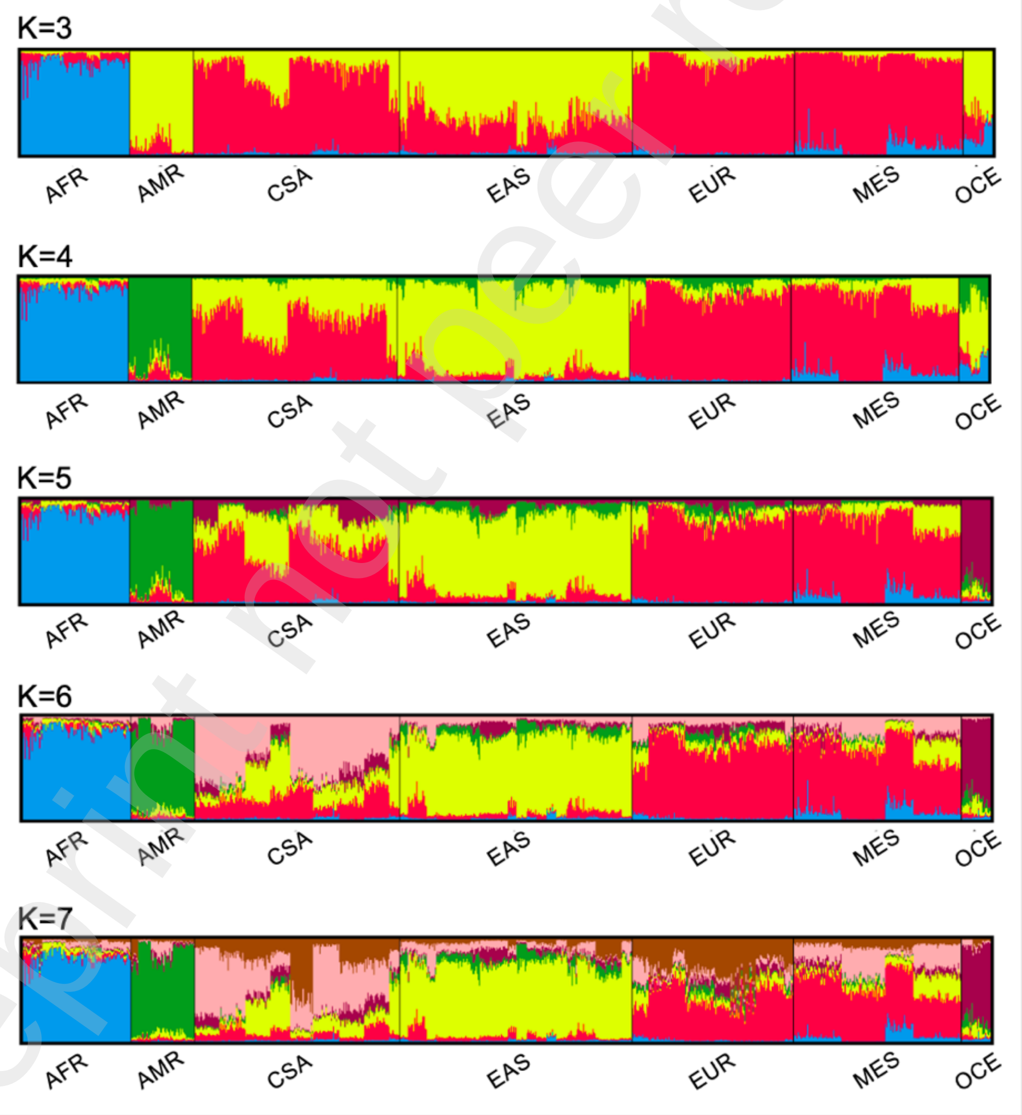

Figure 3. STRUCTURE analysis based on autosomal STR data from the 54 subpopulations of HGDP.

Seven sets of 10 independent runs with the number of clusters ranging from 3 to 7 were conducted. Each bar plot depicts the results from the run with the largest LnP(D) for the given *k*. Penta D and Penta E markers were excluded from this analysis. (AFR: African; CSA: Central South Asia; EAS: East Asia; EUR: European; MES: Middle East; OCE: Oceania).

To verify the distribution of variance in different levels, an AMOVA was performed assuming a hierarchical structure gathering the populations in seven groups: AFR, AMR, CSA, EAS, EUR, MES and OCE without the Penta E and Penta D markers. Most of the variance is observed within populations (95.84%). Differences between the seven groups account for 2.61% of the variance, whereas only 1.55% of the variance occurs due to differences between populations from the same group. An alternative structure, composed of only four groups (merging CSA, EAS, EUR and MES populations in a single group), revealed an increase in the variance between groups: differences between the four groups account for 4.14% of the variance, whereas only 2,17% of the variance occurs due to differences between populations from the same group and as expected most of the variance is observed within populations (93.67%).

The genotypes calculated with HipSTR were compared with a dataset of previously obtained CE-derived genotypes (Algee-Hewitt et al., 2016; Rosenberg et al., 2005). The average number of identical genotypes was 97.44% (median = 99.35%) (Supplementary Table 3), ranging from 88.25% (FGA) to 99.88% (D8S1179). Given the high proportion of genotypes with only one correct allele for some *loci*, these figures are much better when the assignment of correct alleles are taken into account: the average number of correct alleles was 98.49% (median = 99.67%), ranging from 92.12% (FGA) to 99.94% (D8S1179) (Supplementary Table 3). The errors in allele assignment are summarized in Supplementary Table 4. Inconsistencies were considered as “stutter-related” errors when HipSTR failed to detect the smaller allele in situations in which the CE-derived genotype indicated a heterozygote composed of contiguous alleles (e.g., 11/12) and called it as a false homozygous (e.g., 12/12). Stutter-related errors accounted for 13.2% of all errors. Other types of inconsistencies were considered as “stutter-unrelated” errors (86.8%). All 311 errors are detailed in Supplementary Table 5.

Allele frequencies estimated from the HGDP dataset were also compared with STR data retrieved from the same seven major population groups that composed the SPSmart STR browser (Amigo *et al.*, 2009) (Pop.STR) using F_{ST} (Table 5). The Penta E marker presented values <0.05 in all groups except in the Middle East (MES). In general, we observed only 10 significant F_{ST} spread out in four markers: D2S441, D5S818, Penta D, and Penta E.

Table 5. Probabilities obtained by F_{ST} analysis of population differentiation based on genotype frequencies of each STR, comparing population groups from the Human Genome Diversity Project with those from the SPSmart STR browser (Pop.STR). Significant *p*-values ($\alpha = 0.05$) are in boldface. The probabilities that remain significant after the Bonferroni correction for multiple tests ($\alpha_{BONFERRONI} = 0.05/147 = 0.00034$) are also underlined.

Marker	AFR		AMR		CSA		EAS		EUR		MES		OCE	
	F_{ST}	*p*-value	F_{ST}	*p*-value	F_{ST}	*p*-value	F_{ST}	*p*-value	F_{ST}	*p*-value	F_{ST}	*p*-value	F_{ST}	*p*-value
CSF1PO	-0.00915	0.99980+-0.0001	-0.01250	0.85556+-0.0036	-0.00428	0.96010+-0.0021	-0.00433	0.99921+-0.0003	-0.00627	0.99584+-0.0006	-0.00614	0.99871+-0.0004	-0.02731	0.90872+-0.0032
D1S1656	-0.00938	0.99999+-0.0000	-0.01514	0.99994+-0.0000	-0.00492	0.99999+-0.0000	-0.00427	0.99998+-0.0000	-0.00616	0.99999+-0.0000	-0.00593	0.99999+-0.0000	-0.03287	0.99999+-0.0000
D2S441	-0.00905	0.99999+-0.0000	0.18667	**0.00000+-0.0000**	-0.00472	0.99792+-0.0001	-0.00412	0.99736+-0.0002	0.00523	0.11839+-0.0010	0.03174	**0.00063+-0.0001**	-0.03320	0.94017+-0.0008
D2S1338	0.00212	0.27614+-0.0014	-0.01342	0.97511+-0.0005	-0.00061	0.51765+-0.0017	0.00286	0.12064+-0.0010	0.00151	0.26167+-0.0014	-0.00443	0.96563+-0.0006	-0.03561	0.99999+-0.0000
D3S1358	-0.00958	0.99999+-0.0000	-0.01559	0.99697+-0.0002	-0.00480	0.99841+-0.0001	-0.00442	0.99999+-0.0000	-0.00627	0.99912+-0.0001	-0.00605	0.99853+-0.0001	-0.03566	0.99999+-0.0000
D5S818	-0.00902	0.99961+-0.0001	-0.01539	0.99631+-0.0002	-0.00496	0.99965+-0.0001	-0.00428	0.99908+-0.0001	-0.00644	0.99999+-0.0000	0.04888	**0.00000+-0.0000**	-0.03525	0.99999+-0.0000
D7S820	-0.00842	0.99382+-0.0002	-0.01511	0.99693+-0.0002	-0.00487	0.99962+-0.0001	-0.00443	0.99999+-0.0000	-0.00646	0.99999+-0.0000	-0.00570	0.99429+-0.0002	-0.02634	0.92461+-0.0008
D8S1179	-0.00953	0.99999+-0.0000	-0.01601	0.99999+-0.0000	-0.00500	0.99999+-0.0000	-0.00440	0.99999+-0.0000	-0.00632	0.99996+-0.0000	-0.00603	0.99995+-0.0000	-0.03535	0.99999+-0.0000
D10S1248	-0.00920	0.99997+-0.0000	-0.01472	0.99562+-0.0002	-0.00393	0.94223+-0.0007	-0.00385	0.97001+-0.0005	-0.00599	0.99106+-0.0003	-0.00569	0.99402+-0.0002	-0.03438	0.98741+-0.0003
D12S391	-0.00846	0.99995+-0.0000	-0.01439	0.99089+-0.0003	-0.00448	0.99985+-0.0000	-0.00247	0.84891+-0.0011	-0.00443	0.98000+-0.0004	-0.00459	0.98615+-0.0003	-0.03569	0.99999+-0.0000
D13S317	-0.00879	0.99329+-0.0003	-0.01328	0.96635+-0.0006	-0.00487	0.99981+-0.0000	-0.00432	0.99956+-0.0001	-0.00626	0.99988+-0.0000	-0.00605	0.99948+-0.0001	-0.03247	0.99041+-0.0003
D16S539	-0.00948	0.99992+-0.0000	-0.01540	0.99538+-0.0002	-0.00498	0.99999+-0.0000	-0.00418	0.99545+-0.0002	-0.00613	0.99891+-0.0001	-0.00616	0.99991+-0.0000	-0.03353	0.99790+-0.0001
D18S51	-0.00885	0.99994+-0.0000	-0.01508	0.99976+-0.0000	-0.00367	0.97798+-0.0005	-0.00281	0.91754+-0.0009	-0.00611	0.99997+-0.0000	-0.00585	0.99999+-0.0000	-0.01898	0.87383+-0.0010
D19S433	-0.00866	0.99983+-0.0000	-0.01513	0.99990+-0.0000	-0.00453	0.99955+-0.0001	-0.00324	0.93432+-0.0008	-0.00596	0.99906+-0.0001	-0.00596	0.99992+-0.0000	-0.03597	0.99999+-0.0000
D22S1045	0.00985	0.07334+-0.0008	-0.00716	0.55926+-0.0015	-0.00301	0.79675+-0.0013	-0.00370	0.95798+-0.0006	-0.00576	0.98184+-0.0004	-0.00513	0.97348+-0.0005	0.05184	0.06208+-0.0008
FGA	0.00366	0.19120+-0.0013	-0.01013	0.93542+-0.0008	-0.00156	0.67142+-0.0014	-0.00038	0.47693+-0.0016	-0.00438	0.94444+-0.0007	-0.00525	0.99428+-0.0003	-0.02768	0.99177+-0.0003
Penta D	0.02162	**0.00154+-0.0001**	-0.01285	0.94995+-0.0008	-0.00167	0.63971+-0.0015	-0.00248	0.83218+-0.0012	-0.00383	0.86665+-0.0011	-0.00469	0.97591+-0.0005	-0.02859	0.98105+-0.0004
Penta E	0.00953	**0.03937+-0.0006**	0.02947	**0.00501+-0.0002**	0.01702	**0.00010+-0.0000**	0.05375	**0.00000+-0.0000**	0.00603	**0.04632+-0.0007**	0.00332	0.13552+-0.0011	0.04362	**0.01032+-0.0003**
TH01	-0.00952	0.99999+-0.0000	-0.01577	0.99999+-0.0000	-0.00477	0.99722+-0.0002	-0.00437	0.99976+-0.0000	-0.00587	0.98292+-0.0004	-0.00551	0.98326+-0.0004	-0.03620	0.99999+-0.0000
TPOX	-0.00954	0.99987+-0.0000	-0.01540	0.99478+-0.0002	-0.00495	0.99929+-0.0001	-0.00432	0.99626+-0.0002	-0.00514	0.91091+-0.0009	-0.00617	0.99977+-0.0000	-0.03418	0.98061+-0.0004
vWA	-0.00788	0.98409+-0.0004	-0.01544	0.99814+-0.0001	-0.00498	0.99998+-0.0000	-0.00420	0.99766+-0.0001	-0.00590	0.99440+-0.0002	-0.00595	0.99843+-0.0001	-0.03169	0.99310+-0.0003

DISCUSSION

This study offers a STR database from high-coverage next-generation sequencing data derived from the 54 population samples that compose the Human Genome Diversity Project (HGDP).

Accurate STR genotyping from NGS data has been challenging due to the high sequencing error rates and difficulties in aligning repetitive sequences (Fungtammasan et al., 2015). However, Bornman et al. (2012) demonstrated that CODIS *loci* could be accurately called even from complex mixtures using an NGS approach. Notwithstanding that, capillary electrophoresis (CE) is, until now, and will continue to be for a long time, the most used technique to genotype STRs due to its simplicity. CE doesn't offer nucleotide sequence information (Bornman et al., 2012), while an NGS assay allows differentiating isometric alleles (isoalleles), which would permit to increase forensic informativeness (i.e., power of discrimination and power of exclusion) (Hert et al., 2008). In this study, HipSTR was used to differentiate alleles by size, and not by sequence due to the large number of samples processed simultaneously.

HipSTR presented some problems in specific markers like D21S11, Penta D, and Penta E. The D21S11 marker was excluded because HipSTR couldn't genotype it. The same problem has already been reported by Valle-Silva et al. (Valle-Silva et al., 2022) in a previous study. This could be related to the specific region that HipSTR uses to capture this marker, given that D21S11 is a complex marker (Rockenbauer et al., 2014). Moreover, the length of the sequenced alleles may also play a part, given that even the smallest common D21S11 allele (with 26 repeats encompassing 104 nucleotides) is large, and sequencing error rates increase with STR length (Kelkar et al., 2008). Both issues may lead to mapping failure during the alignment step.

On the other hand, we may have failed in genotyping small Penta D alleles that presented less than 5 repeats. This situation mainly affected the African populations: many studies, including the pop.STR data (Amigo et al., 2009), show that Penta D has a very high frequency of the 2.2 allele (0.20%). Also, in this study Penta D presented the lowest successful calling rate (58.56%). Penta E deviated from H-W equilibrium in 27 of 54 populations, being responsible for more than 30% of Hardy-Weinberg departures observed. Because of these problems, Penta D and Penta E were excluded from all interpopulation statistical analyses (Analysis of Molecular Variance, PCoA and clustering analysis). We don't recommend using these markers for population genetics or human identification purposes using the HipSTR software. However, toaSTR (Carsten, 2017; Ganschow et al., 2018; Valle-Silva; et al., 2022) showed very effective Penta D and Penta E genotyping in previous studies. The limitation of this software, which prevented its use in the present study, is that it can only process one sample at a time, while HipSTR can process thousands of samples in parallel.

The D22S1045 marker showed to be monomorphic in an Amerindian population from México, Pima. This population is considered to be composed of descendants of the ancient Hohokam, who have inhabited the Sonoran Desert and Sierra Madre regions for centuries. Today, they are present in two countries, in the USA (Arizona state), as "*The O'odham*", and in Mexico as "*O'ob"* or *"Pima Bajo*" (Schulz et al., 2015). According to the most recent data from the Mexican government, currently, 1.540 Pima exist in the country (HOPE, 2006). The SPSmart STR browser (Amigo et al., 2009) (Pop.STR) revealed precisely the same situation, with only allele 15 being observed. The Pop.STR studied 14 individuals from this population, while HGDP sampled 13 individuals. Small populations typically show a high rate of inbreeding, which produces the fixation of some alleles (Hartl, 2020).

When the genotypes calculated with HipSTR were compared with those from the dataset provided by Algee-Hewitt et al. (2016) and Rosenberg et al. (2005), the average number of identical genotypes was 97.44% (median = 99.35%). The FGA and D22S1045 STRs were the most problematic ones and strongly influenced the average. In the case of FGA, one of the reasons may be the length of some alleles. For instance, although alleles with more than 30 repeats are extremely rare, the largest described FGA allele is composed of 51 tetranucleotide repeats. The observed stutter-unrelated problems could be related to the positioning of flanking regions, tri-allelic patterns or alignment errors. Although it is not reasonable to assign all the inconsistencies to problems in the NGS-based procedure, particularly given that in the two CE-based studies mentioned above were in fact large-scale genome-wide studies that prevented a careful evaluation of each genotype for 1160 STRs in 2034 subjects from worldwide populations, it is noteworthy that HipSTR uses previously obtained bam files. Thus, additional efforts in improving the WGS alignment procedure, particularly considering the repetitive nature of microsatellite regions, may increase the overall accuracy of genotype calling by the HipSTR algorithm. Unfortunately, CE-derived genotypes were unavailable for five markers (D1S1656, D2S1338, D12S391, Penta D, and Penta E), rendering the secondary validation attempt involving the comparison of allele frequencies using pairwise F_{ST} of utmost importance to assess the reliability of their NGS-based genotypes.

The Principal Coordinates Analysis (PCoA) was able to separate the major populations correctly (Figure 1) and also the sub-populations (Figure 2). Similar results were revealed by the clustering analysis (Figure 3). While African, Amerindian and Oceanian populations are clearly differentiated, Asian (CSA, EAS, MES) and European (EUR) populations present high levels of shared ancestry. Although modern humans arose in Africa, the Middle East is considered the cradle of Eurasian civilization (Guest; Sahebkar, 2021), where the world's first civilizations originated. Thanks to its economic supremacy, Europe ended up colonizing the Middle East and leaving a large immigrant community. This situation could be the reason for the genetic similarity between the individuals from these regions. Historically, Central Asia has been an intersection between Western and Eastern Eurasian people, leading to the current high levels of genetic admixture and diversity (González-Ruiz et al., 2012).

The Structure analysis (Figure 3) shows that the African and Native American populations form largely distinct homogeneous clusters, while the Middle Eastern, European, Central, and South Asian populations form a more heterogeneous cluster. These findings reflect the more isolated nature of the former populations and corroborate the idea that although forensic STRs do show relatively low F_{ST}, their high heterozygosities strengthens their capacity to uncover patterns of population clustering, also revealed by other sets of markers (JOBLING, 2022). Our findings agree with the data presented by Pemberton et al. regarding human microsatellite variation on large databases, including the HGDP-CEPH (Pemberton et al., 2013).

Despite the problems already discussed, HipSTR proved to be highly effective for genotyping STR markers from NGS data, mainly for CODIS markers which are the most used in the forensic area. Notwithstanding, we recommend using more than one software to genotype these markers from NGS to obtain high efficiency and circumvent the genotype calling issues we have described.

CONCLUSION

In conclusion, this investigation offers a population genetics perspective based on a comprehensive genotyping analysis of standard STR used in the forensic genetics field concerning the whole Human Genome Diversity Project. Penta D and Penta D Markers were excluded from our analysis because they did not show up as reliable markers. All the remaining genotypes and allele frequencies presented in this study are supported by (a) previous reports that certify HipSTR's reliability, (b) the comparison between CE-derived and NGS-derived genotypes, (c) frequency data reports from worldwide populations, including the large pop.STR database, and (d) the conclusions achieved by our population genetics analysis that corroborates current knowledge regarding modern human demographic history.

FUNDING

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. C.T.M.J. (#312802/2018-8) is supported by a Research Fellowship from CNPq/Brazil.

CONFLICTS OF INTEREST

The authors declare that there are no conflicts of interest.

COMPLIANCE WITH ETHICAL STANDARDS

Not applicable.

REFERENCES

Algee-Hewitt, B. F.; Edge, M. D.; Kim, J.; Li, J. Z. et al (2016). Individual Identifiability Predicts Population Identifiability in Forensic Microsatellite Markers. Curr Biol 26(7):935-942. Doi: 10.1016/j.cub.2016.01.065

Almarri MA, Bergström A, Prado-Martinez J, Yang F., et al (2020). Population Structure, Stratification, and Introgression of Human Structural Variation. Cell. 9;182(1):189-199. Doi: 10.1016/j.cell.2020.05.024.

Amigo J.; Christopher, P.; Toño, S.; Fernandez, F.L., et al (2009). pop.STR—An online population frequency browser for established and new forensic STRs. Forensic Sci. Int. Genet. Suppl. Ser. 2, 361–362. Doi: 10.1016/j.fsigss.2009.08.178

Behjati S, Tarpey PS (2013). What is next generation sequencing? Arch Dis Child Educ Pract 98(6):236-8. Doi: 10.1136/archdischild-2013-304340.

Bergström A, McCarthy SA, Hui R, Almarri MA., et al (2020). Insights into human genetic variation and population history from 929 diverse genomes. Science 20;367(6484). Doi: 10.1126/science.aay5012

Birney E (2021). The International Human Genome Project. Hum Mol Genet. 1,30(R2):R161-R163. Doi: 10.1093/hmg/ddab198.

Bonneville R, Krook MA, Chen HZ, Smith A., et al (2020). Detection of Microsatellite Instability Biomarkers via Next-Generation Sequencing. Methods Mol Biol. 2055:119-132. Doi: 10.1007/978-1-4939-9773-2_5.

Bornman, D.M, Hester, M.E., Schuetter, J.M.; Kasoji, M.D., et al (2012). Short-read, high-throughput sequencing technology for STR genotyping. Biotech. Rapid Dispatches 1–6. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4301848/

Callaway E (2019). First portrait of mysterious Denisovans drawn from DNA. Nature 573(7775):475-476. Doi: 10.1038/d41586-019-02820-0.

Cann HM, de Toma C, Cazes L, Legrand MF., et al (2002) A human genome diversity cell line panel. Science 12;296(5566):261-2. Doi: 10.1126/science.296.5566.261b.

Cavalli-Sforza LL (2005). The Human Genome Diversity Project: past, present and future. Nat Rev Genet. 6(4):333-40. Doi: 10.1038/nrg1596.

Degioanni A, Bonenfant C, Cabut S, Condemi S (2019). Living on the edge: Was demographic weakness the cause of Neanderthal demise? PLoS One 29;14(5):e0216742. Doi: 10.1371/journal.pone.0216742.

Demeter, F.; Zanolli, C.; Westaway, K. E.; Joannes-Boyau, R. et al (2022). A Middle Pleistocene Denisovan molar from the Annamite Chain of northern Laos. Nat Commun, 13(1):2557.

Dodson M, Williamson R (1999). Indigenous peoples and the morality of the Human Genome Diversity Project. J Med Ethics 25(2):204-8. Doi: 10.1136/jme.25.2.204.

Excoffier, L.; Lischer, H.E (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. Doi: 10.1111/j.1755-0998.2010.02847.x

Fan H, Chu JY (2007). A brief review of short tandem repeat mutation. Genomics Proteomics Bioinformatics 5(1):7-14. Doi: 10.1016/S1672-0229(07)60009-6.

Fungtammasan, A.; Ananda, G.; Hile, S.E.; Su, M.S., et al (2015). Accurate typing of short tandem repeats from genome-wide sequencing data and its applications. Genome Res. 25, 736–749. Doi: 10.1101/gr.185892.114.

Ganschow, S.; Silvery, J.; Kalinowski, J.; Tiemann, C (2018). toaSTR: A web application for forensic STR genotyping by massively parallel sequencing. Forensic Sci. Int. Genet. 37, 21–28. Doi: 10.1016/j.fsigen.2018.07.006.

Gettings, K.B.; Ballard, D.; Bodner, M.; Borsuk, L.A.., et al (2019). Report from the STRAND Working Group on the 2019 STR sequence nomenclature meeting. Forensic Sci. Int. Genet. 43, 102165. Doi: 10.1016/j.fsigen.2019.102165.

González-Ruiz M, Santos C, Jordana X, Simón M., et al (2012). Tracing the origin of the east-west population admixture in the Altai region (Central Asia). PLoS One 7(11):e48904. Doi: 10.1371/journal.pone.0048904.

Gouy, A.; Zieger, M (2017). STRAF-A convenient online tool for STR data evaluation in forensic genetics. Forensic Sci. Int. Genet. 30, 148–151. Doi: 10.1016/j.fsigen.2017.07.007.

Guest PC, Sahebkar A (2021). Research in the Middle East into the Health Benefits of Curcumin. Adv Exp Med Biol.1291:1-13. Doi: 10.1007/978-3-030-56153-6_1.

Gymrek, M.; Golan, D.; Rosset, S.; Erlich, Y (2012). lobSTR: A short tandem repeat profiler for personal genomes. Genome Res. 22, 1154–1162. Doi: 10.1101/gr.135780.111.

Halman, A.; Oshlack, A (2020). Accuracy of short tandem repeats genotyping tools in whole exome sequencing data. F1000Res 9, 200. Doi: 10.12688/f1000research.22639.1

Hartl, D (2020). A Primer of Population Genetics and Genomics. 4a ed. Oxford University Press.

Hert DG, Fredlake CP, Barron AE (2008). Advantages and limitations of next-generation sequencing technologies: a comparison of electrophoresis and non-electrophoresis methods. Electrophoresis 29(23):4618-26. Doi: 10.1002/elps.200800456.

Hope, M (2006). Pueblos Indígenas del México Contemporáneo. 2006. Available from: https://www.inpi.gob.mx/2021/dmdocuments/pimas.pdf.

Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009). Inferring weak population structure with the assistance of sample group information. Mol Ecol Resour 9(5):1322-32. Doi: 10.1111/j.1755-0998.2009.02591.x.

Jobling, M.A (2022). Forensic genetics through the lens of Lewontin: Population structure, ancestry and race. Philos. Trans. R. Soc. Lond. B Biol. Sci. 2022, 377, 20200422. Doi: 10.1098/rstb.2020.0422

Kelkar YD, Tyekucheva S, Chiaromonte F, Makova KD (2008). The genome-wide determinants of human and chimpanzee microsatellite evolution. Genome Res 18(1):30-8. Doi: 10.1101/gr.7113408.

Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA., et al (2015). Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. Mol Ecol Resour 15(5):1179-91. doi: 10.1111/1755-0998.12387.

Mallick S, Li H, Lipson M, Mathieson I., et al (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature 13;538(7624):201-206. Doi: 10.1038/nature18964.

Peakall, R.; Smouse, P.E (2012). GenAlEx 6.5: Genetic analysis in Excel. Population genetic software for teaching and research-an update. Bioinformatics 28, 2537–2539. Doi: 10.1093/bioinformatics/bts460

Pemberton TJ, DeGiorgio M, Rosenberg NA (2013). Population structure in a comprehensive genomic data set on human microsatellite variation. G3 Bethesda 20;3(5):891-907. Doi: 10.1534/g3.113.005728

Robinson, J.T.; Thorvaldsdóttir, H.; Wenger, A.M.; Zehir, A., et al (2017). Variant Review with the Integrative Genomics Viewer. Cancer Res. 77, e31–e34. Doi: 10.1158/0008-5472.CAN-17-0337

Rockenbauer E, Hansen S, Mikkelsen M, Børsting C., et al (2014). Characterization of mutations and sequence variants in the D21S11 locus by next generation sequencing. Forensic Sci Int Genet 8(1):68-72. Doi: 10.1016/j.fsigen.2013.06.011.

Rosenberg NA (2006). Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. Ann Hum Genet. 70(Pt 6):841-7. Doi: 10.1111/j.1469-1809.2006.00285.x.

Rosenberg NA, Mahajan S, Ramachandran S, Zhao C., et al (2005). Clines, clusters, and the effect of study design on the inference of human population structure. PLoS Genet 1(6):e70. Doi: 10.1371/journal.pgen.0010070.

Schulz LO, Chaudhari LS (2015). High-Risk Populations: The Pimas of Arizona and Mexico. Curr Obes Rep. 4(1):92-8. Doi: 10.1007/s13679-014-0132-9

Thorvaldsdóttir, H.; Robinson, J.T.; Mesirov, J.P (2013). Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. Brief. Bioinform. 14, 178–192. Doi: 10.1093/bib/bbs017.

Valle-Silva, G.; Frontanilla, T.S.; Ayala, J.; Donadi, E.A., et al (2022). Analysis and comparison of the STR genotypes called with HipSTR, STRait Razor and toaSTR by using next generation sequencing data in a Brazilian population sample. Forensic Sci. Int. Genet. 58, 102676. Doi: 10.1016/j.fsigen.2022.102676

Warshauer, D.H.; Lin, D.; Hari, K.; Jain, R., et al (2013). STRait Razor: A length-based forensic STR allele-calling tool for use with second generation sequencing data. Forensic Sci. Int. Genet. 7, 409–417. Doi: 10.1016/j.fsigen.2013.04.005

Willems, T.; Zielinski, D.; Yuan, J.; Gordon, A., et al (2017). Genome-wide profiling of heritable and de novo STR variations. Nat. Methods 14, 590–592. Doi: 10.1038/nmeth.4267