OPEN-ACCESS STRS DATABASE OF POPULATIONS FROM THE 1000 GENOMES PROJECT USING HIGH COVERAGE PHASE 3 DATA.

Tamara Soledad Frontanilla^{a*}; Guilherme Valle-Silva^{b*}; Jesus Ayala^c; Celso Teixeira Mendes-Junior^a.

**These authors contributed equally to this work*

^aDepartamento 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 Softec S.R.L. Asunción, Paraguay.

*Corresponding author at: 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. Tel.: +55-16-33150417. Fax: +55-16-33154838.
E-mail address: ctmendes@ffclrp.usp.br

ABSTRACT

Accurate STR genotyping from next-generation sequencing (NGS) data has been challenging. Haplotype inference and phasing for STRs (HipSTR) was specifically developed to deal with genotyping errors and obtain reliable STR genotypes from whole-genome sequencing datasets. The objective of this investigation was to perform a comprehensive genotyping analysis of a set of STRs of broad forensic interest from the 1000 Genomes populations and release a reliable open-access STR database to the forensic genetics community. A set of 22 STR markers were analyzed using the CRAM files of the 1000 Genomes Project Phase 3 high-coverage (30x) dataset generated by the New York Genome Center (NYGC). HipSTR was used to call genotypes from 2,504 samples from 26 populations organized into five groups: African, East Asian, European, South Asian, and admixed American. The D21S11 marker could not be detected in the present study. Moreover, the Hardy-Weinberg equilibrium analysis, coupled with a comprehensive analysis of allele frequencies, revealed that HipSTR could not identify longer Penta E (and Penta D at a lesser extent) alleles. This issue is probably due to the limited length of sequencing reads available for genotype calling, resulting in heterozygote deficiency. Notwithstanding that, AMOVA, a clustering analysis using STRUCTURE, and a Principal Coordinates Analysis revealed a clear-cut separation between the four major ancestries sampled by the 1000 Genomes Consortium (AFR, EUR, EAS, SAS). Meanwhile, the AMOVA results corroborated previous reports that most of the variance is (97.12%) observed within populations. This set of analyses revealed that except for larger Penta D and Penta E alleles, allele frequencies and genotypes defined by HipSTR from the 1000 Genomes Project phase 3 data and offered as an open-access database are consistent and highly reliable.

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

INTRODUCTION

Next-generation sequencing (NGS), massively parallel or deep sequencing, is a technology that enables the sequencing of millions of DNA fragments in parallel. NGS can deal with several regions or targets simultaneously, enabling the detection of variation sites or mutations in the genome. This technology has allowed the in-depth study of worldwide human genetic diversity for various purposes, including forensic human identification (1-3).

The 1000 Genomes Consortium is a worldwide collaboration that has produced an extensive catalog of human genetic variation. They sequenced whole genomes from 2,504 individuals from multiple populations derived from five groups: African, East Asian, European, South Asian, and admixed Americans (4). These data are freely available at the International Genome Sample Resource (IGSR) website to generate a variant call format file using a set of specific command lines (5). In 2015, they analyzed the genomes of all individuals using a combination of low-coverage whole-genome sequencing (WGS), deep exome sequencing, and dense microarray genotyping in project's phase 3. They have described worldwide patterns of genomic diversity based on Single Nucleotide Polymorphisms (SNPs), indels, and structural variants (SVs), including deletions, insertions, duplications, inversions, and CNVs (copy-number variants). However, short tandem repeat (STR) markers were not analyzed and studied in depth (6).

STR markers are crucial in human identification due to the high polymorphism levels, being particularly useful for interpreting mixtures of biological samples. In addition to the issue of small-sized amplicons, genotyping these markers using NGS data is particularly difficult due to frequent alignment and stutter errors (7). The accurate genotyping from NGS data has been challenging because of their high sequencing error rates (8). A previous study managed to obtain and analyze STR data from the 1000 Genomes dataset using lobSTR (9). Since high coverage is mandatory for reliable STR genotype calling, a primary concern regarding that study is that the 1000 Genomes data available for lobSTR was generated using shallow sequencing coverage (2x – 6x), so it was severely susceptible to errors.

The New York Genome Center recently resequenced the 2,504 samples from 1000 Genomes Project Phase 3 panel with high (30x) coverage. They have also aligned the sequence data to GRCh38, and these publicly available alignments could be used to call STRs reliably (5, 10).

Some tools have already been developed to analyze STRs markers from NGS data, such as lobSTR (9), STRait Razor (11), toaSTR (12), HipSTR (13), and others. Each software uses different algorithms and flanking regions to capture STR reads.

Haplotype inference and phasing for STRs (HipSTR) was developed for calling microsatellites from Illumina FASTq files. It was designed to deal with genotyping errors and to obtain more robust STR genotypes. HipSTR accomplished this by learning 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 STRs. These factors turn HipSTR into one of the most reliable tools for genotyping STRs from Illumina sequencing data (13).

Different from other tools, it can process hundreds of samples at once. Also, it enables the user to determine the set of STR markers to be analyzed, and the flanking regions used to capture them. In fact, HipSTR showed accurate genotype calling in previous studies. HipSTR accuracy was tested by comparing whole-genome sequencing calls from 118 samples to capillary electrophoresis data, resulting in a 98.8% consistency (13, 14). Recently, we performed a study comparing HipSTR with Strait Razor and toaSTR, and all three methods presented high allele calling accuracy (greater than 97%) (Valle-Silva, Frontanilla, et al., 2021, in preparation). Although data processing with HipSTR is more complex and requires bioinformatics knowledge and some nomenclature adjustments (detailed in Valle-Silva, Frontanilla et al. 2021, in preparation), HipSTR has shown to be the fastest and most appropriate tool to deal with larger datasets, including whole genomes.

This investigation aimed to perform a comprehensive genotyping analysis of a set of STRs of broad forensic interest from the 1000 Genomes populations and release a reliable open-access STR database to contribute to future forensic genetics studies.

METHODOLOGY

Genotype Calling

The CRAM files of 1000 Genomes Project Phase 3 resequenced high-coverage (30x) WGS data generated by the New York Genome Center (NYGC), available at <https://www.internationalgenome.org/data-portal/data-collection/30x-grch38>, were used. The sample size consisted of 2,504 individuals from 26 worldwide populations that compose five population groups: African (AFR), East Asian (EAS), European (EUR), South Asian (SAS), and admixed American (AMR)(4). All samples were sequenced with a paired-end approach (2 x 150pb), using the NovaSeq 6000 Sequencing System (Illumina, Inc.). The high-coverage data provides a more reliable opportunity to genotype STR markers. Thus, we have selected a set of 22 autosomal microsatellites commonly used in forensic practice: 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. These 22 *loci* are autosomal markers that compose the PowerPlex® Fusion System (Promega, Madison, WI).

We ran the HipSTR algorithm for each individual to genotype the 22 STRs based on the human reference genome GRCh38, using a BED file with the flanking regions available in the HipSTR repository (13). We applied the calling filter (15% stutter model) and a minimum of 8 reads to obtain more reliable genotypes. According to a binomial distribution, this minimum number of reads ensures ($p > 0.99$) that a homozygous genotype is called because of lack of variability at a given locus and not because the second allele was not sampled (Valle-Silva, Frontanilla, et al., 2021, submitted).

To perform genotype calling, we used the VCF output file produced by HipSTR taking three parameters into account: the reference allele of each marker, the period (i.e., the length of each STR motif), and the base pair differences (GB) in comparison with the reference allele. Nomenclature adjustments were made for D19S433, Penta D, Penta E, and vWA following (Valle-Silva, Frontanilla, et al. 2021, submitted) recommendations: removal of two repeat units from all D19S433 and vWA alleles, the inclusion of one repeat unit into all Penta D alleles, and removal of two nucleotides from all Penta E alleles. The total number of reads used to genotype each marker in each sample was also recorded.

Statistical Analysis

Allele frequencies, Hardy-Weinberg equilibrium, and forensic parameters [Match Probability (MP), Power of Discrimination (PD), Power of Exclusion (PE), and Polymorphism Information Content (PIC)] were calculated for each population sample or each population group using the GenAlEx (15) and STRAF (16) software.

A Principal Coordinates Analysis (PCoA) using GenAlEx (15), Analysis of Molecular Variance (AMOVA) using Arlequin (17), and clustering analyses using STRUCTURE 2.3.4 (18) were used to explore the distribution of genetic diversity across populations of different ethnic backgrounds. The STRUCTURE analysis was performed for k ranging from 3 to 6, applying the correlated allele frequencies model, 100,000 burn-in steps followed by 100,000 Markov Chain Monte Carlo interactions, in 100 independent runs. The results from the runs with the largest “Estimated Ln Probability of Data” [LnP(D)] were selected and depicted in bar plots created with Distruct 1.1 (19).

We also compared the allele frequencies estimated from the 1000 Genomes dataset with STR data retrieved from the same five major population groups (African, European, East Asian, South Asian, and admixed American) that compose the SPSmart STR browser (Pop.STR) (20). For this purpose, allele

frequencies of each STR for a given population group were compared between the two datasets using F_{ST} and an exact test of population differentiation based on genotype frequencies, using the Arlequin software (17). This comparison was made to verify the reliability of genotype data generated by HipSTR.

RESULTS

STR genotypes defined for each individual from the newest dataset released by the 1000 Genomes Project are available in Supplementary Table 1 as an open-access database. We excluded the D21S11 marker because we did not succeed in genotyping it. Apart from this marker, the mean coverage for calling genotypes ranged from 37.14 (TPOX) to 52.53 (D12S391) (Table 1). Average successful calling rate was 98.59%, ranging from 84.18% (Penta E) to 100% (CSF1PO, D2S441, D2S1338, D3S1358, D5S818, D8S1179, D22S1045, and TPOX) (Table 2).

Table 1. Average coverages obtained for each STR using the HipSTR tool.

Marker	Lowest value	Median	Highest value	Mean	Standard deviation
CSF1PO	21	44	91	44.54	8.29
D1S1656	24	45	92	45.48	8.58
D2S441	26	49	131	49.17	9.04
D2S1338	28	50	105	51.28	9.64
D3S1358	28	51	119	51.73	9.33
D5S818	20	42	98	42.90	8.14
D7S820	20	38	86	39.00	7.78
D8S1179	26	47	96	48.09	8.95
D10S1248	18	40	100	40.75	7.99
D12S391	26	52	113	52.53	9.47
D13S317	11	37	79	37.93	7.50
D16S539	21	44	92	44.70	8.49
D18S51	24	47	91	47.38	9.07
D19S433	19	45	89	45.28	8.62
D22S1045	22	49	111	49.47	9.11
FGA	23	50	118	51.33	9.49
Penta D	19	43	95	43.71	8.74
Penta E	18	41	107	41.49	8.04
TH01	16	40	83	40.73	8.01
TPOX	15	37	86	37.14	7.66
vWA	21	47	105	48.44	9.36

Allele frequencies and forensic parameters estimated for the whole dataset are exposed in Table 2. Allele frequencies and forensic parameters estimated for each of the 26 populations (Supplementary Table 2) and the five population groups (Supplementary Table 3) are available as supplementary data. In general, the most polymorphic *loci* in all populations were D1S1656, D2S1338, D12S391, D18S51, and FGA (Table 2). The analyzed loci were highly informative, with elevated PD values ranging between 86.59% (TPOX) and 97.76% (D1S1656). Combined MP was 5.72×10^{-27} , and combined PE was 0.99999997. Analyzing each locus in each population (Supplementary Table 2), D22S1045 in PEL (71.61%) and D1S1656 in GBR (97.52%) presented the lowest and highest PD, respectively. Combined MP ranged from 1.98×10^{-25} in ACB to 2.20×10^{-21} in PEL.

Adherences of genotype frequencies to Hardy-Weinberg Equilibrium expectations were estimated for each STR at a population level (Table 3). Penta E presented heterozygote deficiency in 24 out of the 26 populations, leading to departures from the Hardy-Weinberg equilibrium. This finding clearly indicates that HipSTR incorrectly called many heterozygous genotypes as homozygous. Disregarding Penta E, the number of deviations ranged from one (D13S317 and D16S539) to five (D19S433 and Penta D), and the number of deviations across populations ranged from zero (ASW and CEU) to seven (PUR), with an average of 2.42 departures in each population. It should be emphasized that if the Bonferroni correction for multiple tests is taken into account, only 39 departures remain significant, most of them (61.53%) concerning Penta E.

Table 2. Allelic frequencies and the forensic parameters estimated for each marker in the whole 1000 Genomes dataset.

Allele	CSF1PO	D1S1656	D2S441	D2S1338	D3S1358	D5S818	D7S820	D8S1179	D10S1248	D12S391	D13S317	D16S539	D18S51	D19S433	D22S1045	FGA	Penta D	Penta E	TH01	TPOX	vWA
5											0.0002						0.0125	0.0861	0.0020		
6							0.0004		0.0002					0.0004			0.0034	0.0017	0.1905	0.0210	
7	0.0198	0.0002				0.0118	0.0178		0.0002		0.0018	0.0004		0.0002			0.0150	0.1084	0.2828	0.0074	
7.3	0.0002																				
8	0.0218	0.0080	0.0006	0.0002		0.0222	0.1842	0.0064	0.0006		0.1394	0.0314		0.0002	0.0006		0.0539	0.0854	0.1287	0.4217	
8.3																			0.0002		
9	0.0334	0.0004	0.0018		0.0004	0.0421	0.0942	0.0066	0.0004		0.0891	0.1805	0.0004	0.0012	0.0002		0.2195	0.0358	0.2314	0.1534	
9.2			0.0052				0.0016													0.0002	
9.3																			0.1491		
10	0.2398	0.0080	0.2264			0.0976	0.2534	0.0953	0.0010		0.0763	0.1123	0.0046	0.0028	0.0154		0.1660	0.0873	0.0150	0.0621	0.0004
10.2			0.0004				0.0004						0.0006								
10.3	0.0002																		0.0002		
11	0.2658	0.0788	0.3401		0.0004	0.3187	0.2506	0.0753	0.0166		0.2650	0.2888	0.0110	0.0210	0.1649		0.1839	0.1774	0.0002	0.2971	0.0014
11.2							0.0004				0.0002			0.0008							
11.3			0.0529																		
11.4																	0.0002				
12	0.3375	0.0732	0.1062		0.0022	0.3097	0.1650	0.1180	0.0684		0.2912	0.2336	0.0805	0.0745	0.0170		0.1506	0.1886		0.0359	0.0008
12.2							0.0002						0.0006	0.0122							
12.3			0.0022											0.0004							
13	0.0693	0.1142	0.0282	0.0002	0.0034	0.1841	0.0275	0.2340	0.2616		0.0982	0.1317	0.1155	0.2706	0.0034		0.1380	0.1065		0.0006	0.0072
13.2													0.0032	0.0341							
13.3			0.0008						0.0002				0.0002								
14	0.0108	0.1472	0.2121	0.0006	0.0777	0.0116	0.0042	0.2418	0.2816	0.0010	0.0370	0.0200	0.1674	0.2726	0.0509		0.0396	0.0669		0.0006	0.1255
14.2													0.0010	0.0623			0.0002				0.0004
14.3		0.0028	0.0004										0.0002								
15	0.0016	0.1772	0.0202	0.0014	0.3075	0.0020		0.1566	0.2220	0.0336	0.0016	0.0014	0.1670	0.1042	0.3313	0.0004	0.0134	0.0380			0.1217
15.1														0.0002							
15.2					0.0006									0.0799							0.0002
15.3		0.0272							0.0002												
16		0.1410	0.0026	0.0318	0.3031	0.0002		0.0559	0.1178		0.0342		0.1414	0.0295	0.2502	0.0006	0.0031	0.0176			0.2251
16.2					0.0002					0.0342	0.0004			0.0240		0.0004					
16.3		0.0560											0.0002								
17		0.0448		0.1328	0.2063			0.0080	0.0268	0.1109			0.1187	0.0046	0.1472	0.0016	0.0007	0.0005			0.2433
17.2		0.0004								0.0014				0.0034							
17.3		0.0796								0.0078											
18		0.0060		0.0897	0.0903			0.0022	0.0022	0.2264			0.0787		0.0170	0.0110					0.1753
18.2										0.0008			0.0002	0.0008			0.0030				
18.3		0.0298								0.0096							0.0002				0.0002
19		0.0006		0.1679	0.0072				0.0002	0.1743			0.0519		0.0014	0.0673					0.0770
19.2										0.0030						0.0016					
19.3		0.0040								0.0044											
20				0.1106	0.0008					0.1415			0.0308		0.0004	0.0906					0.0206
20.2										0.0004			0.0002			0.0012					
20.3		0.0006								0.0002											
21				0.0637						0.0911			0.0130		0.0002	0.1247					0.0006
21.2																0.0034					

22											0.0813						0.0741						0.0076						0.1785						0.0004
22.2																													0.0046						
23											0.1306						0.0552						0.0028						0.1679						
23.2																													0.0034						
23.3																													0.0002						
24											0.1012						0.0180						0.0014						0.1619						
24.2																													0.0048						
25											0.0685						0.0104						0.0004						0.1026						
25.2																													0.0028						
25.3																													0.0002						
26											0.0154						0.0014												0.0436						
26.2																													0.0012						
26.3																													0.0002						
27											0.0030																		0.0137						
27.2																													0.0002						
28											0.0010																		0.0054						
29											0.0002																		0.0026						
30																													0.0002						
N	2504	2500	2504	2504	2504	2504	2494	2504	2500	2502	2485	2502	2497	2496	2504	2490	2235	2108	2502	2504	2499														
Na	11	21	15	18	13	10	13	11	16	22	11	9	27	22	14	31	15	13	10	10	16														
Ho	0.7492	0.8440	0.7380	0.8722	0.7496	0.7220	0.7927	0.8103	0.7536	0.8437	0.7666	0.7838	0.8614	0.8121	0.7364	0.8305	0.7808	0.4877	0.7450	0.6621	0.7943														
He	0.7512	0.8893	0.7729	0.8902	0.7569	0.7567	0.8020	0.8305	0.7836	0.8665	0.8010	0.7983	0.8801	0.8227	0.7755	0.8728	0.8439	0.8803	0.7914	0.7048	0.8226														
MP	0.1059	0.0224	0.0833	0.0226	0.1002	0.0978	0.0690	0.0499	0.0785	0.0324	0.0676	0.0697	0.0267	0.0509	0.0841	0.0287	0.0423	0.0420	0.0737	0.1341	0.0548														
PE	0.5084	0.6831	0.4895	0.7391	0.5091	0.4632	0.5856	0.6183	0.5159	0.6825	0.5386	0.5693	0.7175	0.6217	0.4868	0.6569	0.5638	0.1769	0.5013	0.3722	0.5885														
PD	0.8941	0.9776	0.9167	0.9774	0.8998	0.9022	0.9310	0.9501	0.9215	0.9676	0.9324	0.9303	0.9733	0.9491	0.9159	0.9713	0.9577	0.9580	0.9263	0.8659	0.9452														
PIC	0.7104	0.8790	0.7393	0.8797	0.7169	0.7180	0.7726	0.8088	0.7501	0.8526	0.7738	0.7688	0.8680	0.8021	0.7421	0.8593	0.8244	0.8684	0.7589	0.6575	0.7985														

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.

Table 3. Probabilities of adherence to Hardy-Weinberg equilibrium proportions for each STR in all 26 subpopulations analyzed in the 1000 Genomes Project. Significant p -values ($\alpha = 0.05$) are in boldface. The probabilities that remain significant after the Bonferroni correction for multiple tests ($\alpha_{\text{BONFERRONI}} = 0.05/546 = 0.000092$) are also underlined.

POP	CSF1PO	D1S1656	D2S441	D2S1338	D3S1358	D5S818	D7S820	D8S1179	D10S1248	D12S391	D13S317	D16S539	D18S51	D19S433	D22S1045	FGA	Penta D	Penta E	TH01	TPOX	vWA
ACB	0.8317	0.1003	0.9562	0.6768	0.7289	0.5439	0.0238	0.1626	0.7625	0.9937	0.1509	0.9058	0.1797	0.9795	0.8294	0.8814	0.1985	0.2895	0.4184	0.0678	0.1618
ASW	0.8494	0.8119	0.9805	0.7607	0.8298	0.8945	0.6251	0.9977	0.5225	0.4894	0.2447	0.9927	0.6351	0.3402	0.4035	0.8225	0.3519	0.0724	0.9511	0.9248	0.9409
BEB	0.6165	0.6321	0.1621	0.0216	0.9740	0.6614	0.4515	0.9476	0.8795	0.9124	0.5829	0.6662	0.3397	0.8489	0.6579	0.8494	0.4690	<u>0.0000</u>	0.2040	0.4850	0.9571
CDX	0.9917	0.3823	0.4668	0.9308	0.4053	<u>0.0000</u>	0.5757	0.9797	0.5425	0.4254	0.4678	0.2605	0.6701	0.5218	0.2454	0.1603	0.4560	<u>0.0000</u>	0.0268	0.0025	0.4988
CEU	0.9331	0.5301	0.9948	0.1233	0.1851	0.0674	0.6688	0.2600	0.8395	0.5314	0.8354	0.9559	0.1012	0.9471	0.8354	0.8997	0.5208	<u>0.0000</u>	0.7591	0.4191	0.1209
CHB	0.3105	0.7945	0.0005	0.8305	0.9407	0.8190	0.4159	0.2194	0.7585	0.0003	0.9664	0.4847	0.1581	0.0047	0.9689	0.9747	0.0011	<u>0.0000</u>	0.9424	0.8976	<u>0.0000</u>
CHS	0.2302	0.1077	0.2844	0.6237	0.5740	0.0727	0.9894	0.4883	0.3859	0.6180	0.7626	0.6386	0.3969	0.0391	0.8843	0.9618	0.4278	<u>0.0000</u>	0.1768	0.7723	0.8666
CLM	0.9075	0.3108	0.4415	0.0684	0.3560	0.0470	<u>0.0000</u>	0.8558	0.5450	0.0004	0.1412	0.5892	0.0829	0.9682	0.9999	0.9976	0.6915	<u>0.0000</u>	0.4287	0.4076	0.7427
ESN	0.1175	0.9988	0.9706	0.9750	0.0303	0.2028	0.6773	0.0131	0.8721	0.1069	0.8443	0.9823	0.9529	0.4869	0.0209	1.0000	0.0110	<u>0.0000</u>	0.9238	0.3436	0.0579
FIN	0.9558	0.4612	0.7627	<u>0.0000</u>	0.5922	0.5269	0.8596	0.3818	0.8976	0.9531	0.9688	0.7919	0.2147	0.9665	0.9869	0.7378	0.9930	<u>0.0000</u>	0.9504	0.8369	0.9465
GBR	0.9311	0.9506	0.2788	0.8505	0.9925	0.8037	0.8379	0.9828	0.8791	0.8061	0.2196	0.0259	0.2483	<u>0.0000</u>	0.0317	0.9879	0.2263	<u>0.0000</u>	0.0512	0.4530	0.4718
GIH	0.6965	0.0239	0.8370	0.6288	0.9993	0.4325	0.6899	0.0011	0.8836	0.9808	0.3863	0.6818	0.9979	0.7126	0.4995	0.7371	0.6790	<u>0.0000</u>	0.0770	0.7827	0.3344
GWD	0.1950	0.6832	0.9970	0.9978	0.5107	0.8942	0.2703	0.3213	0.9718	0.9527	0.4987	0.7810	0.0098	0.9973	0.2150	0.1932	<u>0.0000</u>	<u>0.0000</u>	0.1804	0.8530	0.9779
IBS	<u>0.0000</u>	0.8111	0.9373	0.9690	0.1246	0.9874	0.8882	0.8501	0.2448	0.5344	0.9012	0.1202	0.9943	0.9706	0.9373	0.5506	0.0333	<u>0.0000</u>	0.4393	0.6766	0.3111
ITU	0.6156	0.3367	0.0097	0.7459	0.7367	0.9723	0.8685	0.8721	0.1101	0.9617	0.7853	0.5636	0.9777	0.9016	0.9803	0.0404	0.8992	<u>0.0000</u>	0.9027	0.3481	0.4022
JPT	0.8554	0.7945	0.5003	0.7191	0.6566	0.6312	0.7590	0.2612	0.8154	0.7891	0.3862	0.9589	0.9922	0.9952	0.0006	0.7052	0.9159	<u>0.0000</u>	0.7771	0.0002	0.8666
KWV	0.9942	0.0025	0.8299	0.9795	0.9899	0.2980	0.2296	0.3737	0.7030	0.9483	0.5815	0.9489	0.1698	0.5107	0.9073	0.1862	<u>0.0000</u>	<u>0.0000</u>	0.5244	0.0006	0.4226
LWK	0.3621	0.9913	0.9838	0.3363	0.5245	0.8976	0.6751	0.6144	0.8436	0.1478	0.6934	0.7081	0.2947	0.4416	0.9396	0.9998	0.6011	<u>0.0000</u>	0.9548	0.0325	0.9572
MSL	0.8099	0.2244	0.4496	0.7258	0.9316	0.0628	0.0372	0.7088	0.0992	0.9442	0.9897	0.8470	1.0000	0.9398	0.2779	0.1057	0.2086	<u>0.0000</u>	0.0171	0.5908	<u>0.0000</u>
MXL	0.0047	0.0109	0.8297	0.6202	0.1254	0.8631	0.5456	0.0125	0.0509	0.4395	0.8989	0.7887	<u>0.0000</u>	0.3860	0.9233	0.1562	0.5281	<u>0.0000</u>	0.3747	0.4029	0.5094
PEL	0.8460	0.1247	0.9976	0.7122	0.1681	0.9781	0.7676	0.7192	0.9635	0.9402	0.1281	0.6186	0.9833	0.9176	0.8786	<u>0.0000</u>	0.5730	<u>0.0000</u>	0.8399	0.2170	0.0467
PJL	0.8683	0.0073	1.0000	0.6486	0.8966	0.0001	0.8996	0.5188	0.8722	0.5388	0.9943	0.4565	0.9976	0.6490	0.9166	0.0053	0.1485	<u>0.0000</u>	0.0721	0.6604	0.2985
PUR	0.7847	0.0819	0.0058	0.9097	0.1398	0.8342	0.7698	0.4704	0.0034	0.0045	0.0337	0.0556	0.7141	0.0006	0.9810	0.7965	0.7547	<u>0.0000</u>	<u>0.0000</u>	0.2571	0.0191
STU	0.9028	0.5930	<u>0.0000</u>	0.8290	0.2546	0.2661	0.0071	0.4770	0.9882	0.2049	0.0952	0.1927	0.4262	0.2335	0.0001	0.6382	0.2129	<u>0.0000</u>	0.6082	0.5311	0.9627
TSI	0.6299	0.4393	0.9777	0.0805	0.5719	0.4424	0.6240	0.4107	0.0356	0.6573	0.5777	0.9581	0.0074	<u>0.0000</u>	0.1240	0.5698	0.3556	<u>0.0000</u>	0.0203	0.3932	0.8569
YIR	0.6908	0.2395	0.8925	0.3004	<u>0.0000</u>	0.8611	0.7701	0.9277	0.6050	0.3079	0.5197	0.4061	0.9867	0.8417	0.9451	0.9643	0.4923	<u>0.0000</u>	0.4285	0.1299	0.9942

The Principal Coordinates Analysis (PCoA) reveals four different population clusters (Figure 1). The first coordinate separated the cluster of African (AFR) populations on the right side. On the left side, it is possible to observe three different groups: the European (EUR) populations in the upper part, the East Asian (EAS) populations in the lower section, and the South Asian (SAS) populations between them. The CLM, MXL, PEL, and PUR admixed populations cluster with the European populations, while ACB and ASW cluster with Africans, reflecting their ancestry compositions.

Similar results were obtained with the STRUCTURE analysis. Figure 2 depicts STRUCTURE results from runs obtained with k ranging from 3 to 6. With $k = 4$, each cluster reflects one of the major ancestries from the 1000 Genomes Project. Moreover, each of the six admixed American populations presents varying levels of ancestries from the four biogeographical groups.

To verify the distribution of variance in different levels, an AMOVA was performed assuming a hierarchical structure gathering the populations in four groups: AFR, EAS, EUR, and SAS. The six AMR populations were not taken into account. Most of the variance is observed within populations (97.12%). Differences between the four groups account for 2.54% of the variance, whereas only 0.34% of the variance occurs due to differences between populations from the same group.

Figure 1- Principal Coordinates Analysis (PCoA) based on autosomal STR data from the 26 populations from the 1000 Genomes Project. Each point represents a population sample. More details on these populations are available in Supplementary Table 2. Coordinates 1 and 2 account for 39.15% and 19.25 % of the variance, respectively.

K=3

K=4.

K=5

K=6

K=3

K=4.

K=5

K=6

Figure 2- STRUCTURE analysis based on autosomal STR data from the 26 populations from the 1000 Genomes Project. Five sets of 100 independent runs, with the number of clusters ranging from 3 to 6, were conducted. Each bar plot depicts the results from the run with the largest $\text{LnP}(D)$ for the given k .

Allele frequencies estimated from the 1000 Genomes dataset were also compared with STR data retrieved from for the same five major population groups (African, European, East Asian, South Asian, and admixed American) that composed the SPSmart STR browser (Pop.STR) (20) using F_{ST} (Table 4). While AMR (four), EAS (three), EUR (eight), and SAS (four) population groups presented small numbers of markers with significantly different frequencies between the two datasets, AFR presented 17 significant differences. This pattern may be reflecting the set of populations that compose the compared groups. Penta E was the only marker that showed significantly different F_{ST} values in all comparisons. Leaving AFR and Penta E aside, only 15 significant differences out of 80 comparisons are observed: the mean number of statistically significant differences was 0.75 per marker, ranging from zero (eight STRs) to three (D2S441). It should be emphasized that if the Bonferroni correction for multiple tests is taken into account, only 3 of these 15 F_{ST} values remain significant, while 6 out of 16 significant differences observed for AFR (leaving Penta E aside) and all 5 Penta E differences remain significantly different.

Table 4. Probabilities obtained by F_{ST} of population differentiation comparing population groups from the 1000 Genomes Project with those from the SPSmart STR browser (Pop.STR) for each STR. Significant p -values ($\alpha = 0.05$) are in boldface. The probabilities that remain significant after the Bonferroni correction for multiple tests ($\alpha_{\text{BONFERRONI}} = 0.05/105 = 0.00048$) are also underlined.

Marker	AFR		AMR		EAS		EUR		SAS	
	F_{ST}	p -value	F_{ST}	p -value	F_{ST}	p -value	F_{ST}	p -value	F_{ST}	p -value
CSF1PO	0.0084	0.0057±0.0007	-0.00104	0.7571±0.0047	-0.0014	0.6656±0.0053	0.0004	0.2445±0.0044	0.0019	0.1951±0.0036
D1S1656	0.0024	0.0355±0.0019	0.00028	0.3149±0.0046	-0.0024	0.9774±0.0015	0.0045	0.0000±0.0000	0.0021	0.1152±0.0032
D2S441	0.0093	0.0005±0.0002	0.00370	0.0216±0.0015	0.0025	0.1251±0.0027	0.0074	0.0000±0.0000	0.0122	0.0058±0.0007
D2S1338	0.0124	<u>0.0000±0.0000</u>	-0.00048	0.6527±0.0049	-0.0026	0.9940±0.0007	0.0031	0.0008±0.0003	0.0024	0.1044±0.0030
D3S1358	0.0044	0.021±0.0014	-0.00122	0.8369±0.0040	0.0010	0.2578±0.0044	0.0013	0.0835±0.0026	0.0033	0.1143±0.0031
D5S818	0.0030	0.0477±0.0024	0.00609	0.0051±0.0008	-0.0014	0.6936±0.0048	-0.0007	0.7638±0.0047	-0.0014	0.6250±0.0050
D7S820	0.0047	0.0095±0.0009	-0.00105	0.7982±0.0041	-0.0006	0.5151±0.0050	-0.0003	0.5677±0.0054	-0.0023	0.8731±0.0034
D8S1179	0.0033	0.0308±0.0018	-0.00158	0.9817±0.0014	-0.0016	0.7986±0.0042	0.0024	0.0150±0.0012	-0.0003	0.4615±0.0049
D10S1248	0.0017	0.1081±0.0033	-0.00132	0.8984±0.0029	-0.0031	0.9976±0.0005	0.0003	0.2627±0.0044	-0.0001	0.4055±0.0046
D12S391	0.0007	0.1920±0.0041	0.00077	0.1805±0.0036	-0.0019	0.8774±0.0030	0.0008	0.0946±0.0030	-0.0009	0.6155±0.0047
D13S317	0.0160	<u>0.0000±0.0000</u>	0.00017	0.3478±0.0044	-0.0018	0.7993±0.0045	0.0003	0.2904±0.0047	0.0051	0.0354±0.0019
D16S539	0.0105	<u>0.0002±0.0001</u>	-0.00051	0.5773±0.0047	-0.0010	0.5939±0.0046	0.0047	0.0009±0.0003	-0.0010	0.5987±0.0052
D18S51	0.0018	0.0483±0.0019	-0.00138	0.9831±0.0013	-0.0007	0.5795±0.0055	0.0005	0.2028±0.0042	-0.0003	0.4671±0.0047
D19S433	0.0093	<u>0.0003±0.0002</u>	0.00151	0.1085±0.0027	0.0010	0.2476±0.0041	0.0000	0.3759±0.0049	0.0036	0.0660±0.0028
D22S1045	0.0004	0.2943±0.0053	0.00217	0.0766±0.0025	0.0094	0.0055±0.0007	0.0034	0.0087±0.0009	0.0054	0.0516±0.0021
FGA	0.0011	0.1503±0.0037	-0.00089	0.8382±0.0034	0.0036	0.0397±0.0020	0.0007	0.1306±0.0035	0.0018	0.1518±0.0039
Penta D	0.0275	<u>0.0000±0.0000</u>	0.00318	0.0113±0.0010	-0.0015	0.7660±0.0042	0.0003	0.2969±0.0042	0.0000	0.4117±0.0048
Penta E	0.0139	<u>0.0000±0.0000</u>	0.00733	<u>0.0000±0.0000</u>	0.0412	<u>0.0000±0.0000</u>	0.0083	<u>0.0000±0.0000</u>	0.0179	<u>0.0000±0.0000</u>
TH01	0.0103	<u>0.0004±0.0002</u>	0.00090	0.2083±0.0036	-0.0025	0.9262±0.0027	0.0057	<u>0.0001±0.0001</u>	0.0026	0.1425±0.0038
TPOX	0.0039	0.0269±0.0019	-0.00057	0.5622±0.005	-0.0011	0.5341±0.0046	0.0015	0.0939±0.0033	0.0063	0.0455±0.0020
vWA	0.0065	0.0007±0.0003	-0.00095	0.7575±0.0047	0.0005	0.3135±0.0047	-0.0001	0.4312±0.0051	-0.0010	0.5977±0.0049

DISCUSSION

The present study provides the most diverse database of forensic autosomal STR markers from global populations. STRs display high levels of polymorphism, making them attractive for forensic purposes and population genetics studies.

Despite their usefulness in forensic genetics, STRs genotyping from NGS data, especially whole-genome sequencing assays, may be challenging. The main issues concern the high coverage required to accurate genotyping, difficulties to detect longer alleles due to reads of limited sizes, and mutations in flanking regions leading to null alleles (21). These and other problems were previously addressed by Gaag et al. (22) and Valle-Silva et al. (Valle-Silva, Frontanilla, et al., 2021, in preparation).

However, NGS technology allows the simultaneous analysis of dozens of STRs together with different classes of markers that provide complementary contributions for population genetics and human identification. For example, it is not feasible to include SNPs used as predictors of ancestry and phenotypic characteristics into commercial kits that employ capillary electrophoresis; however, it is possible to combine them with STRs in an NGS assay (1).

Despite the challenges addressed here, several studies demonstrated that it is possible to genotype STRs using dedicated bioinformatics tools. Software like LobSTR (9), toaSTR (12), STRait Razor (11), HipSTR (13), and others have shown consistent and accurate results (14)(Valle-Silva, Frontanilla, et al., 2021, in preparation). Moreover, Bornman et al. (8) demonstrated that CODIS *loci* could be accurately called even from mixtures using an NGS approach.

Willems et al. (23) presented an analysis of human STR variation using lobSTR. They used the 1000 Genomes Project phase 1 data generated using low sequencing coverage, which is excessively error-prone. In fact, they reported difficulties in detecting both alleles in each sample, resulting in an overall deficit of heterozygotes.

As previously addressed, several reasons led us to choose HipSTR to call STR genotypes from the 1000 Genomes dataset. By allowing the customization of the flanking regions, it is possible to evaluate almost any STR in hundreds of samples at once with HipSTR. It may appear more complex at first glance, but it is the most appropriate tool to deal with whole genomes. In addition, a recent evaluation of the performance of this software revealed high efficiency and accuracy levels (Valle-Silva, Frontanilla, et al., 2021, in preparation).

Despite the flexibility provided by HipSTR, the D21S11 marker could not be detected in the present study, probably because of the limited length of the Illumina NGS reads (150 bp paired-end reads) (5, 24). Gaag et al. (22), using Illumina sequencing technology, showed that longer alleles from Penta D, Penta E, and FGA presented sequencing errors at the end of the reads, resulting in null alleles and genotyping errors. Using 300 nucleotide-long paired-end reads, we

successfully genotyped this marker with HipSTR (Valle-Silva, Frontanilla, et al., 2021, in preparation), which suggests a sequencing methodology issue rather than a bioinformatics one.

In this study, Penta D and Penta E showed 10.74% and 15.81% of missing data, respectively. As observed for D21S11, this problem is related to the impossibility of detecting longer alleles due to read-length constraints. Supplementary Table 4 shows the comparison between the allele frequencies estimated in the present study with those from the SPSmart STR browser (Pop.STR) (20) concerning the major groups of populations. The straightforward comparison shows that we were not able to detect alleles larger than 18 in Penta E. This failure directly led to Hardy-Weinberg equilibrium deviations (Table 3) due to deficit of heterozygotes in 24 out of the 26 studied populations. Thus, allele frequencies estimated for Penta E are strongly biased towards increased frequencies of shorter alleles and are of limited applicability (Supplementary Table 4). The probabilities obtained with the F_{ST} analysis (Table 4) support this conclusion: Penta E presented significant F_{ST} values in all five comparisons. Although Penta D and FGA also displayed this problem, their undetected alleles usually show low frequencies (Supplementary Table 4). Therefore, this technical issue did not influence the Hardy-Weinberg equilibrium and F_{ST} analysis as much as Penta E. Although this comparison is valid and helpful, it is essential to emphasize that the compared samples correspond to distinct population groups. The African population group in pop.STR comprises mainly East African Somalian individuals (404 out of 507 samples), while the 1000 Genomes Project samples corresponded to West Africa. Similarly, more than 50% of the European population group from pop.STR is composed mainly of U.S. Europeans (1443 out of 2135) (5,20). Taken together, these results support the robustness of the bioinformatics analysis performed in the present study and the reliability of allele frequencies distributions for all loci except for Penta E.

The most polymorphic *loci* in the whole 1000 Genomes dataset were D1S1656, D2S1338, D12S391, D18S51, and FGA. All these markers presented high degrees of polymorphism throughout the world. AMOVA revealed that most of the variance (97.12%) in allele frequencies is observed within populations, corroborating previous studies (25, 26). However, as expected, the principal component analysis (Figure 1), the clustering analysis performed with STRUCTURE (Figure 2), and the AMOVA confirmed that the four ancestral populations groups (AFR, EUR, EAS, SAS) defined by the 1000 Genomes Consortium are indeed significantly different from each other. Since the admixed American populations present different ancestry compositions (Figure 2), most of them cluster with Europeans, while ACB and ASW cluster with Africans (Figure 1).

The results obtained with the STRUCTURE software corroborate the relationship between the different population groups and provide additional

support for the reliability of the genotypes calculated. If $k = 3$, SAS resembles an admixture between EAS and EUR. A specific cluster for SAS emerges with $k = 4$. With $k = 5$, a minor Eurasian (shared between EUR and EAS) component arises. With $k = 6$, the SAS shared ancestry with EUR and EAS becomes more evident. Regarding the admixed American populations, irrespective of the number of clusters considered, ACB and ASW reveal their preeminent African origin, CLM and PUR reveal more extensive European ancestry, and MXL and PEL reveal almost equal amounts of European and Amerindian (i.e., EAS) ancestries. These results fully corroborate the distribution of the populations into the PCoA (Figure 1). Additional clusters do not provide increased resolution with straightforward meaning (data not shown).

In conclusion, this investigation offers a reliable open-access STR database from the 1000 Genomes Project Phase 3 high-coverage (30x) WGS data generated by the NYGC. However, the limited length of sequencing reads introduced noticeable bias in allele frequencies estimated for Penta D and Penta E. The reliability of this dataset is supported (a) by previous studies that attested HipSTR efficiency, (b) by the Hardy-Weinberg equilibrium analysis, (c) by the set of analyses employed to evaluate interpopulation genetic diversity population, and (d) by the comparison of the allele frequencies obtained here with those of other initiatives that used capillary electrophoresis. Despite our eagerness that this open-access database will be of great interest for future forensic population genetics studies, it is imperative to highlight that the current 1000 Genomes Project dataset does not describe the worldwide human genetic diversity. In fact, many biogeographical regions, mainly from Oceania and the Americas, were not sampled, indicating that additional large-scale initiatives may provide further insight into worldwide STR population diversity.

ACKNOWLEDGEMENTS

We thank Dr. Thomas Willems for sharing his knowledge and expertise in bioinformatics and population genetics and his support with the HipSTR tool. 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.

REFERENCES

1. Børsting C, Morling N. Next generation sequencing and its applications in forensic genetics. *Forensic Sci Int Genet*. 2015;18:78-89.
2. Alvarez-Cubero MJ, Saiz M, Martínez-García B, Sayalero SM, Entrala C, Lorente JA, et al. Next generation sequencing: an application in forensic sciences? *Ann Hum Biol*. 2017;44(7):581-92.
3. Ballard D, Winkler-Galicki J, Wesolý J. Massive parallel sequencing in forensics: advantages, issues, technicalities, and prospects. *Int J Legal Med*. 2020;134(4):1291-303.

4. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74.
5. Clarke L, Fairley S, Zheng-Bradley X, Streeter I, Perry E, Lowy E, et al. The international Genome sample resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes Project data. *Nucleic Acids Res*. 2017;45(D1):D854-9.
6. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. *Nature*. 2015;526(7571):75-81.
7. Fungtammasan A, Ananda G, Hile SE, Su MS, Sun C, Harris R, et al. Accurate typing of short tandem repeats from genome-wide sequencing data and its applications. *Genome Res*. 2015;25(5):736-49.
8. Bornman DM, Hester ME, Schuetter JM, Kasoji MD, Minard-Smith A, Barden CA, et al. Short-read, high-throughput sequencing technology for STR genotyping. *Biotech Rapid Dispatches*. 2012;2012:1-6.
9. Gymrek M, Golan D, Rosset S, Erlich Y. lobSTR: A short tandem repeat profiler for personal genomes. *Genome Res*. 2012;22(6):1154-62.
10. Fairley S, Lowy-Gallego E, Perry E, Flicek P. The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. *Nucleic Acids Res*. 2020;48(D1):D941-7.
11. Warshauer DH, Lin D, Hari K, Jain R, Davis C, Larue B, et al. STRait Razor: a length-based forensic STR allele-calling tool for use with second generation sequencing data. *Forensic Sci Int Genet*. 2013;7(4):409-17.
12. Ganschow S, Silvery J, Kalinowski J, Tiemann C. toaSTR: A web application for forensic STR genotyping by massively parallel sequencing. *Forensic Sci Int Genet*. 2018;37:21-8.
13. Willems T, Zielinski D, Yuan J, Gordon A, Gymrek M, Erlich Y. Genome-wide profiling of heritable and de novo STR variations. *Nat Methods*. 2017;14(6):590-2.
14. Halman A, Oshlack A. Accuracy of short tandem repeats genotyping tools in whole exome sequencing data. *F1000Res*. 2020;9:200.
15. Peakall R, Smouse P.E. GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research-an update. *Bioinformatics*. 2012;28:2537-2539.
16. Gouy A, Zieger M. STRAF-A convenient online tool for STR data evaluation in forensic genetics. *Forensic Sci Int Genet*. 2017;30:148-51.
17. Excoffier L, Lischer HE. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour*. 2010;10(3):564-7.
18. Hubisz MJ, Falush D, Stephens M, Pritchard JK. Inferring weak population structure with the assistance of sample group information. *Mol Ecol Resour*. 2009;9(5):1322-32.
19. Rosenberg NA. Distruct: a program for the graphical display of population structure. *Molecular Ecology Notes*, 2004;4:137-138.
20. Jorge A, Christopher P, Toño S, Fernandez FL, Ángel C, Maviky L. *pop.STR*—An online population frequency browser for established and new forensic STRs. *Forensic Sci Int Genet Supplement Series*. 2009;2(1):361-362.
21. Aalbers SE, Hipp MJ, Kennedy SR, Weir BS. Analyzing population structure for forensic STR markers in next generation sequencing data. *Forensic Sci Int Genet*. 2020;49:102364.

22. van der Gaag KJ, de Leeuw RH, Hoogenboom J, Patel J, Storts DR, Laros JFJ, et al. Massively parallel sequencing of short tandem repeats-Population data and mixture analysis results for the PowerSeq™ system. *Forensic Sci Int Genet.* 2016;24:86-96.
23. Willems T, Gymrek M, Highnam G, Mittelman D, Erlich Y, Consortium GP. The landscape of human STR variation. *Genome Res.* 2014;24(11):1894-904.
24. Belsare S, Levy-Sakin M, Mostovoy Y, Durinck S, Chaudhuri S, Xiao M, et al. Evaluating the quality of the 1000 genomes project data. *BMC Genomics.* 2019;20(1):620.
25. Rosenberg NA. A population-genetic perspective on the similarities and differences among worldwide human populations. *Hum Biol.* 2011;83(6):659-84.
26. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, et al. Genetic structure of human populations. *Science.* 2002;298(5602):2381-5.