Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsigen

Short communication

Autosomal STR and SNP characterization of populations from the Northeastern Peruvian Andes with the ForenSeq™ DNA Signature Prep Kit

Evelyn K. Guevara [a,*], Jukka U. Palo [a,b], Jonathan L. King [c], Magdalena M. Buś [c,d], Sonia Guillén [e], Bruce Budowle [c,d], Antti Sajantila [a,f,*]

[a] Department of Forensic Medicine, University of Helsinki, PO Box 40, FI-00014 Helsinki, Finland
[b] Forensic Genetics Unit, Finnish Institute for Health and Welfare, PO BOX 30, FI-00271 Helsinki, Finland
[c] Center for Human Identification, University of North Texas Health Science Center, 3500 Camp Bowie Blvd., Fort Worth, TX 76107, USA
[d] Department of Microbiology, Immunology and Genetics, Graduate School of Biomedical Sciences, University of North Texas Health Science Center, 3500 Camp Bowie Blvd., Fort Worth, TX 76107, USA
[e] Centro Mallqui, San Isidro, Lima 27, Peru
[f] Forensic Medicine Unit, Finnish Institute for Health and Welfare, PO Box 30, FI-00271 Helsinki, Finland

ARTICLE INFO

Keywords:
Autosomal STRs
Autosomal SNPs
Northeastern Peru
Forenseq™ Signature Prep Kit

ABSTRACT

Autosomal DNA data from Peru for human identity testing purposes are scarce in the scientific literature, which hinders obtaining an appropriate portrait of the genetic variation of the resident populations. In this study we genetically characterize five populations from the Northeastern Peruvian Andes (Chachapoyas, Awajún, Wampís, Huancas and Cajamarca). Autosomal short tandem repeat (aSTR) and identity informative single nucleotide polymorphism (iiSNP) data from a total of 233 unrelated individuals are provided, and forensic genetic parameters are calculated for each population and for the combined set Northeastern Peruvian Andes. After correction for multiple testing in the whole dataset of the Northeastern Peruvian Andes, the only departure from Hardy-Weinberg equilibrium was observed in locus rs2111980. Twenty one out of 27 aSTR loci exhibited an increased number of alleles due to sequence variation in the repeat motif and flanking regions. For iiSNPs 33% of the loci displayed flanking region variation. The combined random match probability (RMP), assuming independence of all loci (aSTRs and iiSNPs), in the Chachapoyas, the population with the largest samples size ($N = 172$), was 8.14×10^{-62} for length-based data while for sequence-based was 4.15×10^{-67}. In the merged dataset (Northeastern Peruvian Andes; $N = 233$), the combined RMP when including all markers were 2.96×10^{-61} (length-based) and 3.21×10^{-66} (sequence-based). These new data help to fill up some of the gaps in the genetic canvas of South America and provide essential length- and sequence-based background information for other forensic genetic studies in Peru.

1. Introduction

Over the last decade, massive parallel sequencing (MPS) has facilitated an in-depth genetic exploration of the structures and histories of both ancient and modern Native American populations. Some studies have addressed questions on population dynamics during the initial colonization of the Americas [1,2], while others have focused on issues related to demographic phenomena during pre- and post-colonial times [3]. The level of resolution in these and other studies has revealed hidden population structure in different parts of the Americas [4], as well as levels of diversity and ancestral origins previously overlooked [5]. Though high-resolution data from these genome-wide screenings are valuable in the field of population genetics, they lack compatibility with the markers commonly used by the forensic genetics community which routinely generates a large amount of human population data. In forensic DNA-casework, a relatively small set of standardized STR markers (autosomal, Y-chromosome, X-chromosome) is mostly used.

Although autosomal STR data from many Latin American populations exist and continue to increase yearly (e.g. [6–9]), they are mostly restricted to samples genotyped with capillary electrophoresis (CE), which only provides length-based variation. The use of MPS technologies is however gaining more importance worldwide [10–14] due to its ability to resolve sequence-based variation, to multiplex several marker systems (autosomal STRs and SNPs, Y-chromosome and

* Correspondence to: Department of Forensic Medicine, University of Helsinki, Finland.
E-mail addresses: evelyn.guevara@helsinki.fi (E.K. Guevara), antti.sajantila@helsinki.fi (A. Sajantila).

https://doi.org/10.1016/j.fsigen.2021.102487
Received 21 August 2020; Received in revised form 16 February 2021; Accepted 20 February 2021
Available online 23 February 2021

X-chromosome), genotype a larger number of markers and resolve challenging samples. Despite these advantages, the forensic genetics community has not yet adopted fully this technology and the task of assigning a standardized nomenclature compatible with CE profiles is still underway [15,16]. For these reasons, published MPS data from Latin American populations, excluding metadata such as the highly admixed Hispanics from the USA, are still extremely rare, e.g. [11,17,18]. In Peru, with few exceptions [19,20], published autosomal STR reference data are scarce and available mostly from large cosmopolitan areas (e.g. Lima, Trujillo). These data are typically provided for no more than 15 of the forensically relevant autosomal STR markers included in commercial genotyping kits, and restricted solely to the electrophoretic analyses of allelic length variation.

Here we report the first characterization of length- and sequence-based alleles and their respective frequencies for 27 autosomal short tandem repeats (aSTRs) and 94 identity informative single nucleotide polymorphisms (iiSNPs) obtained with the ForenSeq™ DNA Signature Prep Kit from a sample set of 233 unrelated individuals from the Northeastern Peruvian Andes. Despite the small sample sizes of some of the populations included in this dataset (i.e. Wampís, Huancas, Cajamarca), forensic genetic parameters were estimated in all of them, as MPS data obtained from a smaller sample has been shown to provide a good representation of the allele frequencies from an even larger population in South America [18]. In addition to these, forensic genetic parameters were also calculated for the combined set from the Northeastern Peruvian Andes. To gain a better insight of the possible genetic subdivision in Peru, we characterize the genetic differentiation between our rural Northeastern Peruvian Andes set and that of the admixed cosmopolitan populations described in [19,21].

2. Materials and methods

This study followed ethical guidelines and standards (Helsinki declaration and subsequent amendments) and was approved by permit #329/13/03/00/13, as described in [22]. The sampled populations included the modern Chachapoya people from the northeastern Peruvian montane forests ($N = 172$); two ethnolinguistically affiliated Jivaroan populations from the Amazonian rainforests, Awajún ($N = 25$) and Wampís ($N = 13$); and Huancas ($N = 9$) and the Cajamarca ($N = 14$) from the north Peruvian Andes. All samples were collected from towns and villages across the Amazonas administrative region situated in the eastern slopes of the Northeastern Peruvian Andes. The modern Chachapoya people occupy mostly the montane cloud forests of the Amazonas region where settlements associated to the ancient Chachapoya culture still remain. The ancient Chachapoyas flourished around 900 CE (Common Era) until conquered by the Inca (1475 CE) and Spanish (1532 CE), which gradually led to the dissolution of their culture and language [23]. The Awajún and Wampís, who belong to the Jivaroan (Chicham) ethnolinguistic family, inhabit the tropical rainforests situated north of the Chachapoya territory. Historically, Jivaroan populations fiercely resisted both Inca and Spanish incursions. Although they did not lose their culture or independence, their population suffered a significant reduction during the mid-Colonial period [24,25]. Huancas is a small town located in the heart of the Chachapoya territory, whose ancestors arrived to this region from the central Andes at the time of the Incas [26]. The Cajamarca set consists of immigrants from the neighboring Andean administrative region of Cajamarca who settled in the Chachapoya area during the last five decades. The ancient cultural developments in the Cajamarca region culminated also with the incorporation of most of this area and their populations to the Inca domain [27]. In our study set, the Awajún and Wampís, are the only ones who have retained their indigenous language, whereas the other populations are mostly Spanish-speaking. The criteria of inclusion in each population were 1) the place of birth up to the grandparents' generation in one of the regions/areas described above and 2) surnames either of native or Spanish origin characteristic from those regions/areas.

2.1. Laboratory work

DNA extraction is outlined in [22]. Autosomal genotyping was carried out using the ForenSeq™ DNA Signature Prep Kit (Verogen; San Diego, Ca., USA) following the protocol described in [11].

2.2. Data analysis

Output FASTQ files from the ForenSeq™ Universal Analysis Software (UAS) pipeline were processed with STRait Razor v2s [28] to obtain the length- and sequence-based genotypes for all samples. Additionally, 16 reference sets from Peru ($\leq$ 16 STRs, altogether 1813 samples) [19,21] were included in downstream analyses. Also, for the calculation of genetic distances (F_{ST}) with a more complete set of markers (aSTRs and iiSNPs), five additional populations were included [10,11,29].

Determination of STR allele frequencies in each study population, tests for departures from Hardy-Weinberg Equilibrium (HWE), detection of linkage disequilibrium (LD) and F_{ST} estimates were performed in Arlequin ver. 3.11 [30]. Correction for multiple testing was accomplished with the Benjamini-Hochberg method [31]. Pairwise F_{ST} values were visualized with a Neighbor-Joining (NJ) tree in MEGA 7 [32].

Population genetic parameters relevant for forensic genetics, such as expected heterozygosity (H_{exp}), power of discrimination (PD), polymorphic information content (PIC), random match probabilities (RMP), power of exclusion (PE) and typical paternity index (TPI) were calculated for aSTRs and iiSNPs on a per marker basis using an in-house Excel-based workbook developed by one of the authors (JLK) and with the tool STRAF [33]. All calculations were performed independently for all study populations and for the pooled set labeled Northeast Peruvian Andes.

3. Results

After data curation and filtering, 233 autosomal genotypes (27 aSTRs and 94 iiSNPs) from the combined northeastern Andean Peruvian populations were included in the analyses. Complete length- and sequence-based autosomal STR genotypes (excluding locus Penta E) obtained for 208 samples were also submitted to STRidER (STRs for Identity ENFSI Reference Database; https://strider.online) [34] for quality control check (STRidER dataset reference STR000340).

3.1. Summary statistics

In the sequence-based STR data, the most polymorphic loci varied according to population (Table S1a). In general, D12S391 showed the highest allele number, ranging from 33 in the Chachapoyas (Fig. 1) to 8 in Huancas. In fact, D12S931 was the most polymorphic locus in all populations except in the Cajamarca and Huancas, where the highest number of alleles was observed in D1S1656 and in D2S1338. The number of observed alleles ranged from 6 (D10S1248, D22S1045,

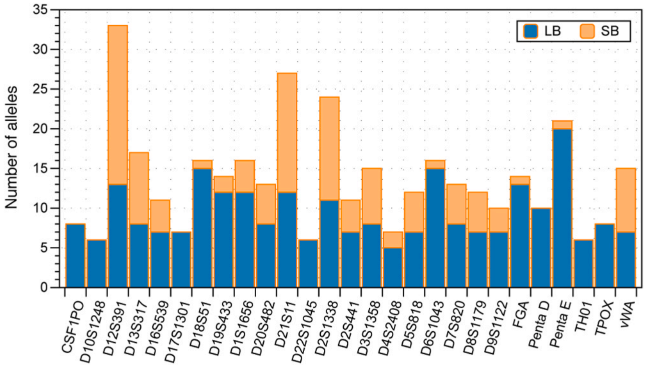

Fig. 1. Length-based (LB) and sequence-based (SB) allele counts for 27 aSTRs in the Chachapoyas.

TH01) to 33 (D12S391) in the Chachapoyas, from 4 (CSF1PO, TPOX) to 18 (D12S391) in the Awajún, from 3 (CSF1PO, D17S1301, D4S2408) to 12 (D12S391) in the Wampís, from 3 (D4S2408, TH01) to 9 (D1S1656, D2S1338) in Huancas, and from 3 (TPOX) to 12 (D1S1656) in Cajamarca.

Length-based alleles showed different patterns of variation, with the highest allele number observed for Penta E in Chachapoyas (Fig. 1), D12S391 in the Awajún, D6S1043 in the Wampís, D1S1656 in the Huancas and three loci in the Cajamarca (D1S1656, D2S1338 and Penta E). The number of observed alleles ranged from 5 (D4S2408) to 20 (Penta E) in the Chachapoyas, from 4 (CSF1PO, D3S1358, D4S2408, D9S1122 and TPOX) to 12 (D12S291) in the Awajún, from 2 (D4S2408) to 11 (D6S1043) in the Wampís, from 3 (D22S1045, D4S2408 and TH01) to 9 (D1S1656) in Huancas, and from 3 (D20S482, D9S1122 and TPOX) to 10 (D1S1656, D2S1338 and Penta E) in Cajamarca.

For a detailed summary of the number of STR alleles refer to Table S1a. Six STR loci did not show any increase when comparing length- and sequence-based genotypes across all populations, namely CSF1PO, D10S1248, D17S1301, D22S1045, TH01 and TPOX. In contrast, loci D12S391, D13S317, D21S11, D2S1338, D3S1358 and vWA showed the greatest allele gain for example in the Chachapoyas.

In the case of iiSNPs, 31 out of 94 showed an increase in the number of alleles when including sequence variation in the flanking region adjacent to the target SNP (Table S1c). From these, 14 were considered adventitious microhaplotypes for this study (i.e., observing three or more alleles at a locus above 1% in the Chachapoyas). As these amplicons were designed to target a single polymorphism (i.e., iiSNPs listed in the documentation of the kit), any additional information obtained from the flanking region may be considered fortuitous. As such, most loci with increased allele variation consisted largely of low-frequency variants (e.g., amplicon rs9905977). Similar to previous observations [35,36], the amplicons for rs9905977 and rs1109037 had the most observed alleles, with the latter amplicon producing four alleles with frequencies above 0.20 in the Chachapoyas. Three additional loci (rs10776839, rs2830795, and rs876724) had three or more alleles with frequencies above 10% (the latter with three above 20%) suggesting that these markers may be particularly useful for mixture analysis given the small size of the amplicons.

Length- and sequence-based allele frequencies are provided in Table S1b and Table S1c, respectively. For more detailed information on sequence variation in the repeat (STRs) and flanking regions, as well as chromosomal coordinates for reported regions for each locus, refer to Supplementary Table S2. All aSTR sequence variants were also submitted to STRSeq [16], of which 31 novel sequences are being accessioned (indicated in Table S2, column STRSeqQueue).

When examining the study populations independently, a significant HWE departure was observed only in the Chachapoyas at one locus (D1S1656) after Benjamini-Hochberg correction (Table S3). Additionally, after correction for multiple testing, the pooled set labeled Northeastern Peruvian Andes, had only one marker (rs2111980) deviating from HWE. Finally, when using 27 STRs for the pooled metadata from Peru (study and reference populations), two loci showed significant departures (Penta E and D3S1358).

Linkage disequilibrium was assessed for 27 aSTRs and 94 iiSNPs in each study population and in the combined set Northeastern Peruvian Andes (Table S4). After correction for multiple testing (Benjamini-Hochberg), two pairs of syntenic loci showed departures from expectations. One in the Chachapoyas (rs13182883 - rs338882), and one pair in the Cajamarca (rs6955448 - rs917118). From these two, only the pair rs6955448 - rs917118 has a separation distance of less than 0.3 cM (~0.2 cM), which is considered the limit of physical linkage. In the merged set Northeastern Peruvian Andes, one syntenic loci showed deviation after correction for multiple testing (rs13182883-rs338882); however, the distance between the start and end coordinates of these loci exceeds 0.3 cM.

When including the full set of markers (115 markers with no missing data), F_{ST} pairwise values were small ($F_{ST} \leq 0.03$, $p \leq 0.05$) among our study populations and the Yavapai from North America as well as with the Hispanic reference population (HIS) from the USA (Table S1d) [10, 29]. These affinities are reflected in the NJ tree where these populations cluster together (Fig. 2).

3.2. Forensic genetic parameters

In all populations and the combined Northeastern Peruvian dataset, H_{exp} as well as RMP, PD, PIC, PE and TPI were calculated for length- and sequence-based allele data (Table S5). Intra-allelic STR sequence variation was observed in the majority of loci, which increased their H_{exp}. The percentage increases ranged from 0.2% (Penta E in Chachapoyas) to 65.7% (D4S2408 in the Wampís). The average increase in H_{exp} over all STR loci in the combined Northeastern Peruvian Andes dataset was 5.0%.

Due to the small sample sizes ($N < 30$) of several of our study populations (i.e. Awajún,Wampís, Huancas, Cajamarca), the forensic parameters results are presented and discussed only for the Chachapoyas and for the combined Northeastern Peruvian Andes dataset. When using length-based genotypes for STRs, the average single-locus RMP in the Chachapoyas was 0.123 ± 0.084, whereas for sequence-based genotypes it was 0.096 ± 0.069. In the case of iiSNPs, the average single-locus was 0.452 ± 0.112 while when including sequence flanking region variation, it was 0.436 ± 0.123. Very similar values were obtained for the merged Northeastern Peruvian Andes set: length-based STR data RMP = 0.126 ± 0.085 and sequence-based data RMP = 0.101 ± 0.070. For iiSNPs, the single-locus RMP was 0.452 ± 0.115 while the sequence-based flanking region variation RMP was 0.436 ± 0.125.

When considering only aSTRs in the Chachapoyas, the combined RMP for length-based data was 2.35×10^{-28} while for sequence-based it was 1.18×10^{-31}. In the case of iiSNPs, the combined RMP were 3.46×10^{-34} and 3.50×10^{-36} (including the flanking region). The RMP for length- and sequence-based variation in the Chachapoyas, including all markers, were 8.14×10^{-62} and 4.15×10^{-67}, respectively. The PD was higher for the combined set of aSTRS and iiSNPs for both length- and sequence-based genotypes. When PD values were calculated separately for aSTRs and iiSNPs in the Chachapoyas, the latter provided higher values (Table S5). Similar to other measures of diversity (i.e. H_{exp}) and as expected, PIC is greater for aSTRs than for iiSNPs, irrespective of length- and sequence-based genotypes. Most aSTRs in the Chachapoyas are highly informative (PIC > 0.5) with the exception of D4S2408 (0.4) for length-based data. PE values are highest for the full set of markers followed by iiSNPs and finally aSTRs. The PE probability was lowest for length-based data in the Chachapoyas. The combined TPI is higher for aSTRs than for the full set of markers for both length- and sequence-based data, and this is heightened when examining sequence-based variation (Table S5).

In the merged Northeastern Peruvian Andes set, the combined aSTRs RMP for length- and sequence-based data were 9.85×10^{-28} and 9.64×10^{-31}, respectively. For iiSNPs, the combined RMP was 3×10^{-34} and when including flanking region sequence variation it was $3.34 \times 10^{-}$

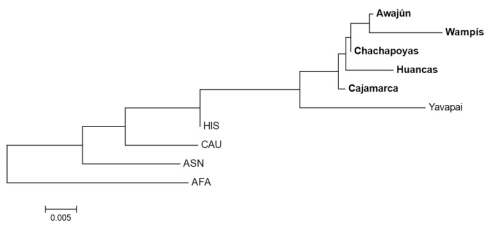

Fig. 2. NJ tree with all populations. The study populations cluster with the Yavapai of North America [11]. AFA = African American, ASN = Asian American, CAU = Caucasian, HIS = Hispanic [10,29], all from the USA. A scale bar of 0.005 indicates the genetic distance (F_{ST}) between populations.

[36]. The RMP when incorporating all markers were 2.96 × 10^{-61} (length-based data) and 3.21 × 10^{-66} (sequence-based data). The PD was higher for the combined set of markers (aSTRS and iiSNPs) for both length- and sequence-based (flanking region for iiSNPs) genotypes. When independent PD calculations for aSTRs and iiSNPs were performed, higher values were observed in the SNP set (Table S5). Length- and sequence-based PIC values are greater for aSTRs when compared to iiSNPs or the full set of markers. Similar to the Chachapoyas, most aSTRs in this Northeastern Peruvian Andes combined set are highly informative (PIC > 0.5) with the exception of D4S2408 (0.282). PE values are highest for the full set of markers followed by iiSNPs and finally aSTRs. Again, here the PE probability was lowest for length-based data. The combined TPI is higher for aSTRs than for the full set of markers (for both length- and sequence-based data), which becomes even higher when examining sequence-based variation (Table S5).

4. Discussion

Even though assessments of the allele frequency variation and forensic genetic parameters were calculated for all populations, we focus on the Chachapoyas ($N = 172$) and on the merged dataset labeled Northeastern Peruvian Andes ($N = 233$) in most of the discussion.

All individual study populations and also the combined Northeastern Peruvian Andes set can be assumed to be in HWE. The significant single deviations observed in the Chachapoyas and the pooled set can probably be attributed to genotyping errors since substructure, or any other population-level cause, would lead to more systematic deviations across loci. The HWE in the merged dataset, together with the largely non-significant differentiation between populations, suggests that the allele frequencies obtained from the pooled dataset could be used for estimating the power of evidence in forensic casework. On the other hand, the two deviations observed in the merged Peruvian metadata (study and reference populations) for STRs may indicate underlying population substructure within Peru since sample sets from different geographic origins were pooled (Andes, Amazon, Coast). After correction for multiple testing (Benjamini-Hochberg), two pairs of syntenic loci showed significant LD deviations, one pair in the Chachapoyas (rs13182883 - rs338882) and one pair in the Cajamarca (rs6955448 - rs917118). It is known that genomic regions separated at >0.3 cM in most parts of the human genome are rarely physically linked [37], such as in the case of the deviation in the Chachapoyas. However, the LD deviation in the Cajamarca involves a pair separated by ~0.2 cM, suggesting that some population-level mechanism may be at play in this particular population. Nevertheless, in case of a population level cause, LD deviations should be more systematic, i.e. involving a larger number of locus pairs, which suggests that this observation is due to chance alone. For the Northeastern Peruvian Andes set, the same syntenic pair of loci as in the Chachapoyas showed deviation after correction for multiple testing (rs13182883 - rs338882); however, like in the previous case, this may be negligible as it occurred only in one locus pair separated by larger distances at which LD is unlikely to occur.

In the majority of STR loci, there was an increase in the number of alleles when comparing length- and sequence-based alleles in all populations. Only six out of 27 (22%) aSTRs did not show any increase (CSF1PO, D10S1248, D17S1301, D22S1045, TH01 and TPOX), a finding in line with previous observations [13]. Despite their small sample sizes, the other study populations showed also substantial allelic variation for different sets of loci (Table S1a).

When examining length-based versus sequence-based genotypes in the Chachapoyas and in the Northeastern Peruvian Andes set, an increase in expected and observed heterozygosity was observed for the majority of aSTR loci, which adds to the growing body of evidence of length and sequence-diversity differences already observed for other populations in the Americas [10,12,36]. In the case of iiSNPs, about 33% of them exhibited flanking region variation (sequence-based alleles), 14 of which were microhaplotypes. Microhaplotypes have various advantages over the commonly used gold standards, STRs, e.g. absence of stutter peaks, alleles of the same size at a given locus and low mutation rates [38,39]. These traits allow them to be successfully used for various forensic genetic applications such as human identification, ancestry prediction, kinship, detection and deconvolution of DNA mixtures as well as treatment of degraded DNA samples [35,40–42]. Although information from worldwide populations has been examined for targeted microhaplotypes [38], a more comprehensive characterization of variation for these adventitious microhaplotypes in the Americas is largely missing as there are few studies that have reported flanking-region variation of the iiSNPs targeted in the ForenSeq™ DNA Signature Prep Kit [35,36].

Similar to previous studies [13], the combination of aSTR and iiSNP loci information provides more statistical power in the calculation of forensic genetic parameters such as RMP. In the Chachapoyas, the dataset with the largest sample size ($N = 172$), some loci showed a large decrease while others exhibited little or no decrease (Table S6). For both length- and sequence-based data in the Chachapoyas and in the Northeast Peruvian Andes sets, PD was higher when combining aSTRs and iiSNPs. Considering these two marker types separately, iiSNPs provided better resolution. PIC values are more informative for aSTR data (PIC > 0.5) in both population sets, with the exception of D4S2408 (and this only for length-based data). The patterns of PD and TPI in the Chachapoyas are mirrored in the Northeastern Peruvian Andes set for both length- and sequence-based data. The highest PE was observed for the full set of markers, followed by iiSNPs and then aSTRs. The combined TPI is higher for aSTRs than in the case of the full set of markers, which becomes larger when considering sequence-based data (Table S6).

A phylogenetic tree shows that the Hispanic dataset (HIS) situates close to the European (CAU) suggesting the Hispanic set may have higher levels of admixture than our study populations, which is also supported by a cluster formed only by Native American populations positioned more distant to datasets of different continental origins. A similar pattern has also been observed with model-based clustering methods for these datasets [43].

5. Conclusion

Here we report the first characterization of length- and sequence-based genotypes from five populations ($N = 233$) from Northeast Peru with the ForenSeq™ DNA Signature Prep Kit. There were no systematic HWE and LD deviations (syntenic loci) for individual populations or in the combined set from the Northeastern Peruvian Andes, which suggests this pooled set can be used for forensic genetic casework. However, when our full aSTR dataset was merged to other populations from Peru, two loci were not in HWE, which implies there is substructure within Peru, since populations from different macro regions were included (Amazon, Andes and Coast).

For aSTRs, sequence-based genotypes showed an increase in allelic variation (21 loci) giving more resolution and power in the calculation of various forensic genetic parameters (e.g. RMP, PD, TPI). Similarly, several iiSNPs (33%) exhibited flanking region variation which also included adventitious microhaplotypes (15%). The new layer of information provided by the sequence-based data reported in this study contributes to uncover underlying genetic variation in populations from Peru, which is also useful for other forensic applications such as kinship or DNA mixtures.

The genotypes, allele frequencies and the estimated forensic genetic parameters, from both length- and sequence-based data in this study, will contribute to increase the available data sets in the Americas and provide essential information for other forensic genetic studies in Peru and South America.

Acknowledgments

We are grateful to all sample donors. We thank Xiangpei Zeng for his guidance during laboratory work. The funding bodies Medicinska Understödsföreningen Liv och Hälsa, The Finnish Society of Sciences and Letters, and Kone Foundation and the Academy of Finland are acknowledged for financial support to Antti Sajantila. Also, we would like to thank Kone Foundation for two grants awarded to Evelyn K. Guevara. The regional Government of Amazonas and the Municipality of Chachapoyas are also acknowledged for providing financial support. The funding bodies did not have any role in the research or in the preparation of this article.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fsigen.2021.102487.

References

[1] D. Reich, N. Patterson, D. Campbell, A. Tandon, S. Mazieres, N. Ray, M.V. Parra, W. Rojas, C. Duque, N. Mesa, L.F. García, O. Triana, S. Blair, A. Maestre, J.C. Dib, C.M. Bravi, G. Bailliet, D. Corach, T. Hünemeier, M.C. Bortolini, F.M. Salzano, M. L. Petzl-Erler, V. Acuña-Alonzo, C. Aguilar-Salinas, S. Canizales-Quinteros, T. Tusié-Luna, L. Riba, M. Rodríguez-Cruz, M. Lopez-Alarcón, R. Coral-Vazquez, T. Canto-Cetina, I. Silva-Zolezzi, J.C. Fernandez-Lopez, A.V. Contreras, G. Jimenez-Sanchez, M.J. Gómez-Vázquez, J. Molina, Á. Carracedo, A. Salas, C. Gallo, G. Poletti, D.B. Witonsky, G. Alkorta-Aranburu, R.I. Sukernik, L. Osipova, S. A. Fedorova, R. Vasquez, M. Villena, C. Moreau, R. Barrantes, D. Pauls, L. Excoffier, G. Bedoya, F. Rothhammer, J.M. Dugoujon, G. Larrouy, W. Klitz, D. Labuda, J. Kidd, K. Kidd, A. Di Rienzo, N.B. Freimer, A.L. Price, A. Ruiz-Linares, Reconstructing Native American population history, Nature 488 (7411) (2012) 370–374.

[2] P. Flegontov, N.E. Altınışık, P. Changmai, N. Rohland, S. Mallick, N. Adamski, D. A. Bolnick, N. Broomandkhoshbacht, F. Candilio, B.J. Culleton, O. Flegontova, T. M. Friesen, C. Jeong, T.K. Harper, D. Keating, D.J. Kennett, A.M. Kim, T. C. Lamnidis, A.M. Lawson, I. Olalde, J. Oppenheimer, B.A. Potter, J. Raff, R. A. Sattler, P. Skoglund, K. Stewardson, E.J. Vajda, S. Vasilyev, E. Veselovskaya, M. G. Hayes, D.H. O'Rourke, J. Krause, R. Pinhasi, D. Reich, S. Schiffels, Palaeo-Eskimo genetic ancestry and the peopling of Chukotka and North America, Nature 570 (7760) (2019) 236–240.

[3] Harris, et al., Evolutionary genomic dynamics of Peruvians before, during, and after the Inca Empire, Proc. Nat. Acad. Sci. 115 (28) (2018) E6526–E6535.

[4] J. Lindo, et al., Ancient individuals from the North American Northwest Coast reveal 10,000 years of regional genetic continuity. 2017. 114(16): p. 4093–4098.

[5] C. Posth, N. Nakatsuka, I. Lazaridis, P. Skoglund, S. Mallick, T.C. Lamnidis, N. Rohland, K. Nägele, N. Adamski, E. Bertolini, N. Broomandkhoshbacht, A. Cooper, B.J. Culleton, T. Ferraz, M. Ferry, A. Furtwängler, W. Haak, K. Harkins, T.K. Harper, T. Hünemeier, A.M. Lawson, B. Llamas, M. Michel, E. Nelson, J. Oppenheimer, N. Patterson, S. Schiffels, J. Sedig, K. Stewardson, S. Talamo, C. C. Wang, J.J. Hublin, M. Hubbe, K. Harvati, A. Nuevo Delaunay, J. Beier, M. Francken, P. Kaulicke, H. Reyes-Centeno, K. Rademaker, W.R. Trask, M. Robinson, S.M. Gutierrez, K.M. Prufer, D.C. Salazar-García, E.N. Chim, L. Müller Plumm Gomes, M.L. Alves, A. Liryo, M. Inglez, R.E. Oliveira, D.V. Bernardo, A. Barioni, V. Wesolowski, N.A. Scheifler, M.A. Rivera, C.R. Plens, P.G. Messineo, L. Figuti, D. Corach, C. Scabuzzo, S. Eggers, P. DeBlasis, M. Reindel, C. Méndez, G. Politis, E. Tomasto-Cagigao, D.J. Kennett, A. Strauss, L. Fehren-Schmitz, J. Krause, D. Reich, Reconstructing the deep population history of central and South America, Cell 175 (5) (2018) 1185–1197.e22.

[6] A. Castillo, A. Pico, A. Gil, L. Gusmão, C. Vargas, Genetic variation of 23 STR loci in a Northeast Colombian population (department of Santander), Forensic Sci. Int. Genet. Suppl. Ser. 7 (1) (2019) 33–35.

[7] J. Aguilar-Velázquez, M. Sthepenson-Ojea, M.D. García-King, H. Rangel-Villalobos, Admixture and population structure in Mayas and Ladinos from Guatemala based on 15 STRs, Forensic Sci. Int. Genet. Suppl. Ser. 7 (1) (2019) 776–777.

[8] L. González-Herrera, J.E. Sosa-Escalante, P. López-González, M.J. López-González, R.Y. Gamboa-Magaña, R.G. Herrera-Diaz, K.A. Piña-Dzul, S.F. León- Acosta, R. I. Flores-Baas, J. Bautista-González, M.R. Rivera-Guzman, H. Rangel-Villalobos, Forensic statistical parameters of 22 autosomal STRs in Mestizos from the Peninsula of Yucatan, Mexico, Forensic Sci. Int. Genet. Suppl. Ser. 7 (1) (2019) 491–493.

[9] A.K. Zambrano, A. Gaviria, M. Vela, C. Rodríguez-Pollit, P. Guevara-Ramírez, A. López-Cortés, I. Armendáriz-Castillo, J.M. García-Cárdenas, S. Guerrero, P. E. Leone, A. Pérez-Villa, V. Yumiceba, G. Fiallos, C. Gruezo, C. Paz-y-Miño, Genetic variation of high-altitude Ecuadorian population using autosomal STR markers, Forensic Sci. Int. Genet. Suppl. Ser. 7 (1) (2019) 62–64.

[10] N.M.M. Novroski, J.L. King, J.D. Churchill, L.H. Seah, B. Budowle, Characterization of genetic sequence variation of 58 STR loci in four major population groups, Forensic Sci. Int. Genet. 25 (2016) 214–226.

[11] F.R. Wendt, J.D. Churchill, N.M.M. Novroski, J.L. King, J. Ng, R.F. Oldt, K. L. McCulloh, J.A. Weise, D.G. Smith, S. Kanthaswamy, B. Budowle, Genetic analysis of the Yavapai Native Americans from West-Central Arizona using the Illumina MiSeq FGxTM forensic genomics system, Forensic Sci. Int. Genet. 24 (2016) 18–23.

[12] K.B. Gettings, L.A. Borsuk, C.R. Steffen, K.M. Kiesler, P.M. Vallone, Sequence-based US population data for 27 autosomal STR loci, Forensic Sci. Int. Genet. 37 (2018) 106–115.

[13] A. Delest, et al., Sequenced-based French population data from 169 unrelated individuals with Verogen's ForenSeq DNA signature prep kit, Forensic Sci. Int. Genet. 47 (2020), 102304.

[14] F. Casals, R. Anglada, N. Bonet, R. Rasal, K.J. van der Gaag, J. Hoogenboom, N. Solé-Morata, D. Comas, F. Calafell, Length and repeat-sequence variation in 58 STRs and 94 SNPs in two Spanish populations, Forensic Sci. Int. Genet. 30 (2017) 66–70.

[15] W. Parson, D. Ballard, B. Budowle, J.M. Butler, K.B. Gettings, P. Gill, L. Gusmão, D. R. Hares, J.A. Irwin, J.L. King, P. Knijff, N. Morling, M. Prinz, P.M. Schneider, C. V. Neste, S. Willuweit, C. Phillips, Massively parallel sequencing of forensic STRs: considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements, Forensic Sci. Int. Genet. 22 (2016) 54–63.

[16] K.B. Gettings, L.A. Borsuk, D. Ballard, M. Bodner, B. Budowle, L. Devesse, J. King, W. Parson, C. Phillips, P.M. Vallone, STRSeq: a catalog of sequence diversity at human identification Short Tandem Repeat loci, Forensic Sci. Int. Genet. 31 (2017) 111–117.

[17] D.S.B.S. Silva, F.R. Sawitzki, M.K.R. Scheible, S.F. Bailey, C.S. Alho, S.A. Faith, Genetic analysis of Southern Brazil subjects using the PowerSeq (TM) AUTO/Y system for short tandem repeat sequencing, Forensic Sci. Int. Genet. 33 (2018) 129–135.

[18] C.G. Bottino, R. Silva, R.S. Moura-Neto, Analysis of 124 SNP loci included in HID Ampliseq identity panel in a small population of Rio de Janeiro, Brazil, Forensic Sci. Int. Genet. Suppl. Ser. 7 (1) (2019) 243–244.

[19] F. Messina, T. Di Corcia, M. Ragazzo, C. Sanchez Mellado, I. Contini, P. Malaspina, B.M. Ciminelli, O. Rickards, C. Jodice, Signs of continental ancestry in urban populations of Peru through autosomal STR loci and mitochondrial DNA typing, Plos One 13 (7) (2018), e0200796.

[20] G.C. Iannacone, et al., Peruvian genetic structure and their impact in the identification of Andean missing persons: a perspective from Ayacucho, Forensic Sci. Int. Genet. Suppl. Ser. 3 (1) (2011) e291–e292.

[21] M. Talledo, M. Gavilan, C. Choque, L. Aiquipa, J. Arévalo, Y. Montoya, Comparative allele distribution at 16 STR loci between the Andean and coastal population from Peru, Forensic Sci. Int. Genet. 4 (4) (2010) E109–E117.

[22] E.K. Guevara, J.U. Palo, S. Guillén, A. Sajantila, MtDNA and Y-chromosomal diversity in the Chachapoya, a population from the northeast Peruvian Andes-Amazon divide, Am. J. Hum. Biol. 28 (2016) 857–867.

[23] I. Schjellerup, Incas y españoles en la conquista de los chachapoya, Fondo Editorial de la Pontificia Universidad Católica del Perú and Instituto Francés de Estudios Andinos, Lima, 2005, p. 641.

[24] M.E. Reeve, Regional interaction in the Western Amazon - the early colonial encounter and the jesuit years 1538-1767, Ethnohistory 41 (1) (1993) 106–138.

[25] A.C. Taylor, The Western Margins of Amazonia from the Early Sixteenth to the Early Nineteenth Century, In The Cambridge History of the Native Peoples of the Americas, in: F. Salomon, S.B. Schwartz (Eds.), Cambridge University Press, 1999, pp. 188–256.

[26] R.P. Colin, *El Pueblo de Huancas.* Boletín de la Sociedad Geográfica de Lima, XVII (XXI) (1907) 465–470.

[27] D.G. Julien, Late Pre-Inkaic Ethnic Groups in Highland Peru: An Archaeological-Ethnohistorical Model of the Political Geography of the Cajamarca Region, in: Latin American Antiquity, 4, 1993, pp. 246–273.

[28] J.L. King, F.R. Wendt, J. Sun, B. Budowle, STRait Razor v2s: Advancing sequence-based STR allele reporting and beyond to other marker systems, Forensic Sci. Int. Genet. 29 (2017) 21–28.

[29] J.D. Churchill, N.M.M. Novroski, J.L. King, L.H. Seah, B. Budowle, Population and performance analyses of four major populations with Illumina's FGx Forensic Genomics System, Forensic Sci. Int. Genet. 30 (2017) 81–92.

[30] L. Excoffier, H.E.L. Lischer, Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows, Mol. Ecol. Resour. 10 (3) (2010) 564–567.

[31] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. R. Stat. Soc. Ser. B Methodol. 57 (1) (1995) 289–300.

[32] S. Kumar, G. Stecher, K. Tamura, MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets, Mol. Biol. Evol. 33 (7) (2016) 1870–1874.

[33] A. Gouy, M. Zieger, STRAF-A convenient online tool for STR data evaluation in forensic genetics, Forensic Sci. Int. Genet. 30 (2017) 148–151.

[34] M. Bodner, I. Bastisch, J.M. Butler, R. Fimmers, P. Gill, L. Gusmão, N. Morling, C. Phillips, M. Prinz, P.M. Schneider, W. Parson, Recommendations of the DNA Commission of the International Society for Forensic Genetics (ISFG) on quality control of autosomal short tandem repeat allele frequency databasing (STRidER), Forensic Sci. Int. Genet. 24 (2016) 97–102.

[35] J.L. King, J.D. Churchill, N.M.M. Novroski, X. Zeng, D.H. Warshauer, L.H. Seah, B. Budowle, Increasing the discrimination power of ancestry- and identity-informative SNP loci within the ForenSeg (TM) DNA Signature Prep Kit, Forensic Sci. Int. Genet. 36 (2018) 60–76.

[36] F.R. Wendt, J.L. King, N.M.M. Novroski, J.D. Churchill, J. Ng, R.F. Oldt, K. L. McCulloh, J.A. Weise, D.G. Smith, S. Kanthaswamy, B. Budowle, Flanking region

variation of ForenSeq (TM) DNA Signature Prep Kit STR and SNP loci in Yavapai Native Americans, Forensic Sci. Int. Genet. 28 (2017) 146–154.

[37] A.R. Rogers, How population growth affects linkage disequilibrium, Genetics 197 (4) (2014) 1329–1341.

[38] K.K. Kidd, W.C. Speed, A.J. Pakstis, D.S. Podini, R. Lagacé, J. Chang, S. Wootton, E. Haigh, U. Soundararajan, Evaluating 130 microhaplotypes across a global set of 83 populations, Forensic Sci. Int. Genet. 29 (2017) 29–37.

[39] F. Oldoni, D. Bader, C. Fantinato, S.C. Wootton, R. Lagacé, R. Hasegawa, J. Chang, K. Kidd, D. Podini, A massively parallel sequencing assay of microhaplotypes for mixture deconvolution, Forensic Sci. Int. Genet. Suppl. Ser. 7 (1) (2019) 522–524.

[40] F. Oldoni, R. Hart, K. Long, K. Maddela, S. Cisana, M. Schanfield, S. Wootton, J. Chang, R. Lagace, R. Hasegawa, K. Kidd, D. Podini, Microhaplotypes for ancestry prediction, Forensic Sci. Int. Genet. Suppl. Ser. 6 (2017) E513–E515.

[41] L. Bennett, et al., Mixture deconvolution by massively parallel sequencing of microhaplotypes, Int. J. Leg. Med. 133 (3) (2019).

[42] C. Turchi, F. Melchionda, M. Pesaresi, P. Fattorini, A. Tagliabracci, Performance of a massive parallel sequencing microhaplotypes assay on degraded DNA, Forensic Sci. Int. Genet. Suppl. Ser. 7 (1) (2019) 782–783.

[43] E.K. Guevara, J.U. Palo, S. Översti, J.L. King, M. Seidel, M. Stoljarova, F.R. Wendt, M.M. Bus, A. Guengerich, W.B. Church, S. Guillén, L. Roewer, B. Budowle, A. Sajantila, Genetic assessment reveals no population substructure and divergent regional and sex-specific histories in the Chachapoyas from northeast Peru, PLOS ONE 15 (12) (2020), e0244497.