A forensic population database in El Salvador: 58 STRs and 94 SNPs

Ferran Casals[1,2], Raquel Rasal[1], Roger Anglada[1], Marc Tormo[1,3], Núria Bonet[1], Nury Rivas[4], Patricia Vásquez[5]*, Francesc Calafell[6]*

[1]Genomics Core Facility, Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Parc de Recerca Biomèdica de Barcelona, 08003 Barcelona, Catalonia, Spain

[2] Departament de Genètica, Microbiologia i Estadística, Universitat de Barcelona, Barcelona, Spain

[3] Scientific IT Core Facility, Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Parc de Recerca Biomèdica de Barcelona, 08003 Barcelona, Catalonia, Spain

[4] Instituto de Medicina Legal Dr. Roberto Masferrer, San Salvador, El Salvador.

[5] Asociación Pro-Búsqueda de Niñas y Niños Desaparecidos de El Salvador, 27 calle Pnte. No.1329 Colonia Layco, San Salvador, El Salvador

[6]Institute of Evolutionary Biology (UPF-CSIC), Department of Experimental and Health Sciences, Universitat Pompeu Fabra, Barcelona, Spain

* Corresponding authors

Abstract

We have genotyped the 58 STRs (27 autosomal, 24 Y-STRs and 7 X-STRs) and 94 autosomal SNPs in Illumina ForenSeq™ Primer Mix A in a sample of 248 men and 143 women from El Salvador, Central America. Regional division (Centro, Oriente, Occidente) showed in almost all cases F_{ST} values not significantly different from 0, and further analyses were applied only to the undivided, country-wide population. The overall random match probability (RMP) decreased from 6.79×10^{-31} in length-based genotypes in the 27 autosomal STRs to 1.47×10^{-34} in repeat-sequence based genotypes. Combining the autosomal loci in this set, RMP reaches 2.97×10^{-70}. In a population genetic analysis, El Salvador showed the lowest F_{ST} values with US Hispanics both for autosomal and X-STRs; however, it was much closer to Native Americans for the latter than for the former, in accordance with the well-known gender-biased admixture that created most Latin American populations.

Keywords: Massive Parallel Sequencing; Repeat Sequence-Based Alleles; Missing persons

Introduction

El Salvador is the smallest nation (~21,000 Km^2) in Central America and the only country in the Central American isthmus that has no coast in the Caribbean. Its population has been estimated for 2019 at 6,704,864 people, 61.7% of whom live in urban areas [1]. Administratively, the country is divided in three regions (Centro, Oriente and Occidente, meaning respectively Center, East, and West), further subdivided into 14 departments. About 76% of the population is concentrated in 5 departments,namely, San Salvador, La Libertad, Santa Ana, Sonsonate, and San Miguel). Three main indigenous groups have inhabited the Salvadoran territory: the Nahua-Pipiles, the Lencas and the Kakawira (Cacaopera) [2]. According to the last census, which was conducted in 2007, 83% of the Salvadoran population self-identify as Mestizo (which is a common term throughout Latin America to describe the product of historic admixture mostly between Native Americans and Europeans), followed by European (15%), Afro-descendant (0.13%) and only 0.23% as Native American, although the latter figure could be an underestimate [3].

El Salvador, like other countries in the region, has suffered repression and human rights violations since colonial times. Social injustice persisted and was a major trigger of the 1980-1992 civil war. In the aftermath of war, poverty and gang violence increased, which were a major factor in inducing emigration, mostly to the USA, but also to other countries. These historical events, as well as the current situation, have generated a need for genetic identification: the Civil War produced large numbers of unidentified casualties and missing persons, both adults and forcibly adopted children. After the peace agreements, and in the absence of a response from the Government, in 1994 the relatives of these missing children formed the Pro-Búsqueda Association (http://www.probusqueda.org.sv/) with the support of the Jesuit priest Jon Cortina; Pro-Búsqueda manages a database of genetic profiles of relatives and young people already found; in addition, in recent years it has begun including young people looking for their families. On the other hand, the number of missing migrants at the U.S.-Mexico border (and which fraction of those are Salvadorans) is unknown. In 2010, government entities, relatives of deceased and missing migrants, and the Argentine forensic anthropology team (EAAF) promoted a forensic data bank of unidentified migrants from El Salvador, which is part of the Proyecto Frontera, a regional mechanism for the forensic exchange between unidentified remains and missing persons along the Central America-Mexico-United States of America migration corridor (https://bancoforenseelsalvador.org/quienes-somos/). Obviously, identification of missing people requires the availability of the allelic frequencies of the genetic markers used in casework. Previous reports of allele or haplotype frequencies in loci of forensic interest include autosomal STRs [4–9], X-STRs [10], Y-STRs [11–14], and mtDNA sequences [15].

Recently, several platforms have become available to apply massive parallel sequencing (MPS) to the genotyping of STRs and SNPs in a forensic context. MPS offers the possibility of multiplexing a much larger number of markers than capillary electrophoresis, and of combining SNPs and STRs. Large numbers of markers assure that, even if results cannot be obtained for part of the markers due to DNA degradation in the sample, still the remaining loci

can provide a likelihood ratio that can be clearly interpreted as indicating or rejecting a match. And a large number of markers may be needed to resolve cases involving distant relatives as references [16], as it is often the case in the identification of missing persons decades after their deaths.

Sequencing rather than sizing STRs allows extracting more information from most of the STRs. In particular, the Verogen Forenseq™ Primer Mix A (Verogen, San Diego, CA) ,which contains 58 STRs and 94 SNPs, together with the with the Universal Analysis Software (UAS) provided by the manufacturer yields for each successful genotype two different types of information: a length-based (LB) genotype, in accordance with the numeric genotype that sizing by capillary electrophoresis would have yielded, and a repeat-sequence based genotype (RSB), that is, the sequence haplotype of the repeat region of each STR (the flanking region sequence is not available through the UAS). A number of studies [17–20], among others, have shown that, while the overall informativeness (as measured by a priori statistics) increases moderately from LB to RSB genotypes, the number of different alleles and the number of rare alleles register more substantial increases.

We hereby report the allele frequencies in 391 samples of El Salvador of the Verogen Forenseq™ Primer Mix A loci, as a resource in the quest for the identification of the missing persons in the country.

Methods

DNA was obtained from either buccal cells or saliva for 402 samples from the general population of El Salvador, after appropriate written informed consent. One sample was excluded because it showed a number of STRs with three and four alleles, probably indicating contamination. Ten additional samples were also removed because they were detected as first- or second-degree relatives of other samples in the database. Thus, the final sample size was 391 individuals (248 men and 143 women), subdivided into 196 samples from the Centro region, 90 from Occidente and 105 from Oriente.

DNA was extracted from a total of 402 samples from buccal cells and saliva in Buccal DNA Collector (Bode Technology, 282 samples) or EasiCollect (Qiagen, 180 samples) collectors. For the extraction we used 8 card punches (1.2mm each) per sample and the PrepFiler BTA kit (Thermofisher Scientific, Waltham MA, USA). Few changes were made from the original protocol (https://assets.thermofisher.com/TFS-Assets/LSG/manuals/cms_099065.pdf) such as the elution buffer volume which was 40ul instead of 50ul and the incubation time that was increased to 20min. Quantifications were performed using 2 ul per sample and the High Sensitivity Qubit method in an Invitrogen Qubit 4 Fluorometer (Thermofisher Scientific). This project was reviewed and approved by the National Committee for Health Research, El Salvador (reference CNEIS/2018/030), on July 24th, 2018.

Samples were sequenced for the Verogen ForenSeq™ Primer Mix A loci according to the manufacturer's protocol. Sample volume for amplification and subsequent library preparation was 5 ml, at a DNA concentration of 0.2 ng/ul. The pooled libraries were sequenced in a 351×31 cycles run with the MiSeq FGx™ instrument (Illumina, San Diego, CA, USA) following the supplier's protocol. We performed six sequencing runs in a standard flow cell, with 22, 88,

96, 96, 94 and 96 samples, and one run in a micro flow cell with 16 samples, plus the manufacturer-supplied positive and negative controls in each run.

STR allele sequences were retrieved from the report generated by the Forenseq UAS interface and inspected by means of an in-house R script (IFator for autosomal STRs, YIFator for Y-STRs, and XIFator for X-STRs, available from github https://github.com/ fcalafell/) [17] . These scripts allow uncovering much more sequence diversity than that reported by the Forenseq UAS interface, which only highlights sequence variants when they are found in isometric heterozygotes, that is, in individuals carrying two alleles of the same length but different sequence. Note that the Forenseq UAS provides exclusively the repeat region sequence, and thus all of our subsequent analyses are based on the repeat region sequence (as processed with our scripts) and cannot include the flanking region. We used a shorthand notation for sequence-based alleles which was based on the called number of repeats plus a lower-case letter indicating an approximation to the repeat structure, consistently with [17]. For instance, STR D3S1358 has the general structure TCTA [TCTG]x [TCTA]y; length is given by 1 + x + y, which we supplement with *a* if x = 1, *b* if x = 2, *c* if x = 3 or *d* if x = 4. Thus, allele TCTA $[TCTG]_1$ $[TCTA]_{13}$ is denoted 15a. See [17] for further details. The full list of RSB variants and their notation can be found in Supplementary Table 1. Allele length information was taken directly from the Forenseq UAS. Hardy-Weinberg equilibrium (HWE) and F_{ST}, were computed with Arlequin 3.5 [21]. We computed two a priori informativeness statistics: power of discrimination [22] and the chance of excluding a putative father in a paternity trio if the mother is known [23], by direct calculation from allele frequencies using MS Excel. Information in the F_{ST} distance matrix was extracted and plotted by means of multidimensional scaling (MDS), which was computed with the isoMDS function in the MASS library in R. Y haplogroups (i.e., the main branches of the Y-SNP tree) were predicted from Y-STR haplotypes using the nevgen (http://nevgen.org) Bayesian predictor, and adapting the nomenclature of the resulting haplogroups to that suggested by [24] and used in http://yhrd.org. For SNP analyses, the analytical threshold (e.g., the lower limit of detection) was set to 0 and the interpretation (allele calling) threshold was set to 2.5% [17], to avoid false negative calls.

Results

Allele frequency and forensic informativeness data were generated for 27 autosomal STRs, 7 X-STRs, 24 Y-STRs and 94 SNPs in a sample of 248 men and 143 women from the general population of El Salvador. Samples were divided according to the region of origin within the country (Centro, Occidente and Oriente). However, after applying the Bonferroni correction for multiple testing taking into account the number of loci in each category (autosomal STRs, X-STRs, Y-STRs, and SNPs), F_{ST} values among the regions were not statistically significantly different from 0 ($p>0.05$) for all but one locus (DYS390, F_{ST} =0.0514) (see Supplementary Tables 2-5, and 9), and thus, for all subsequent analyses, the El Salvador sample has been treated as a single entity.

Allele frequencies, heterozygosities, Hardy-Weinberg p-values and a priori informativeness statistics for 27 autosomal STRs in a population sample of 391 individuals from El Salvador are presented in Supplementary Tables 2 and 3, respectively for LB and RSB alleles. Sample sizes for these analyses ranged from 488 to 782, with an average of 758 and a median of 758 chromosomes. We could generate genotypes for all individuals in the sample for 18 of the 27 STRs. In five additional loci, missing genotypes were less than 5% of the total population sample. The most problematic loci were PentaE, with 37.6% missing genotypes, and PentaD, with 24.0%. LB genotypes were in Hardy-Weinberg equilibrium ($p > 0.05$) at all but four STRs (D22S1045, D5S818, PentaD, and PentaE) after Bonferroni correction. Random match probability (RMP) was 6.79×10^{-31}, which was of the same order of magnitude of that in Catalans [17] or US Hispanics [18]. The joint chance of paternity exclusion was $1-(2.8\times10^{-11})$. RSB variation was detected in 20 out of 27 autosomal STRs, for a total of 490 RSB alleles, up from 293 LB alleles. Exactly the same four STRs mentioned above (namely, D22S1045, D5S818, PentaD, and PentaE) also failed HWE after Bonferroni correction for RSB genotypes. RMP decreased to 1.47×10^{-34}, that is 4,629 times lower that the LB-based RMP. Rare alleles (defined here arbitrarily as those with a frequency <1%) can have an important contribution to solving cases involving distant relatives [16]. We found 83 LB rare alleles with 178 (45.5%) individuals carrying at least one, and up to four rare alleles across all autosomal STRs. These figures clearly increased for RSB: there were a total of 237 RSB rare alleles, with 306 (78.3%) individuals carrying at least one rare allele.

The LB allele frequencies for the autosomal STRs were subjected to the quality controls (QC) implemented in STRidER (http://strider.info)[25,26] and received report number STR000381. Since STRider requires complete profiles, we did not deposit in STRideER the four STRs with >5% missing genotypes (namely, D5S818, D22S1045, PentaD, and PentaE), and, subsequently, neither the remaining 19 (out of 391) samples still carrying missing genotypes. Note that the Forenseq UAS provides only the repeat region for most STRs, and only in six cases 10-31 bp of either 3' or 5' flanking region. This makes QC difficult; for example, in D13S317, a 4-bp deletion has been described in the flanking region[25], which implies that the same number of repeats in the repeat region can produce alleles of different length, which would be called differently by capillary electrophoresis. In locus D7S820, the Forenseq UAS provides 16 bp of the 3' flanking sequence; this means that the sequence starts in the middle of an $[A]_8$ repeat tract. Usually, the sequence starts with AAA, but, in a few cases, it is given as AAAA and automatically called as a .1 imperfect allele. The user cannot verify whether indeed an additional A is present in this poly-A tract or whether the extra A is the product of an error in the trimming of the sequence by the Forenseq UAS.

Allele frequencies, heterozygosities, Hardy-Weinberg test results and F_{ST} values for X-STRs are presented in Supplementary Tables 4 and 5 for LB and RSB alleles. A volunteer was a heterozygote for three of the seven X-STRs, yet genotypes were also recovered for 22 out of 24 Y-STRs. Additionally, AMELX and AMELY were sequenced in this individual at 116x and 68x coverages, respectively. Contamination seems to be ruled out by the fact that, in 27 autosomal STRs, 5 show no amplification, in 6 only one allele is detected, and for 16 STRs, two alleles are detected, with no autosomal STR showing more than two alleles. Thus, these results fit the

expected pattern of a XXY karyotype, which also agrees with the fact that the volunteer self-identifies as a male. Even though the sample was comprised of 142 women and 249 men, the maximum number of X chromosomes in the population sample was 534 rather than 533 (=2×142+249). Allele frequencies, expected heterozygosities and F_{ST} values were estimated from total chromosome samples ranging from 323 to 534, with an average 489 and a median 532 chromosomes; Hardy-Weinberg tests were performed on the female samples, which ranged from 96 to 143, with an average of 132 and a median of 143. Genotypes could be generated for all samples at 3 out of 7 loci, and in a fourth locus, only one sample had a missing genotype. On the contrary, missing genotype rates were 6.6% for DXS10135, 12.3% for DXS8378, and 41.9% for DXS10103. Missing rates were higher for men (47.2%) than for women (41.9%) in DXS10103, while the reverse was true for DXS8378 (14.0% in women and 11.3% in men) and DXS10135 (6.9% in women and 6.5% in men). Variation in the repeat sequence was present in three X-STR loci, and, adding up across all seven loci, 82 different LB and 129 RSB alleles were found. Considering that the STRs within the pairs DXS10135-DXS8378, DXS7132-DXS10074, and DXS10103-HPRTB are in close proximity of each other, we also estimated haplotype frequencies by direct counting in males and informative (i.e., heterozygote in at most one locus within a particular pair) females (Supplementary Tables 6 and 7). Sample sizes for these haplotypes were respectively 324, 346, and 237 chromosomes. The availability of repeat sequences implied that the number of different haplotypes increased from 173 LB haplotypes to 227 RSB haplotypes.

Sample sizes for the 24 Y-STRs present in the Verogen ForenSeq™ primer mix A ranged from 134 to 249, with an average of 223 and a median of 241. For seven loci, all individuals could be genotyped, and, in an additional four loci, missing values were <5%. On the contrary, DYS392 had 46.2% missing genotype calls, DYS389II reached 34.1%, and the value for DYS448 was 29.3%. It should also be noted that DYS438 produced genotypes in two individuals, at coverages 52X and 71X, in which no other Y-STR could be genotyped and that were heterozygotes for 7 and 5 out of 7 X-STRs. Presumably, these are XX persons with a spurious amplification of DYS438. The average and median number of missing genotypes per individual were 2.56 and 1, respectively, and, for 76 individuals, we could produce complete haplotypes. We could detect RSB alleles in 11 Y-STRs; overall, the number of alleles increased from 244 LB alleles to 367 RSB alleles. However, RSB variation did not imply an increase in the number of haplotypes (Supplementary Table 8), since the 76 males (out of a population sample of 249 men) for which we could generate complete haplotypes all carried different LB haplotypes. The average F_{ST} value among the three Salvadoran regions was low (0.0021), and it was not significantly different from zero ($p>0.05$) in all but one locus, namely DYS390, with F_{ST}=0.0631 ($p=0.0001$). In Table 1, we report the frequencies of the predicted haplogroups for this subset. In particular, we found haplogroup E1b1a (5.3%), which is much more frequent in African populations than elsewhere, as well as haplogroup Q1a2-M346 (13.2%), which is found almost exclusively in Native Americans. The rest of haplogroups and their frequencies, once the frequencies of these two components are subtracted, are similar to those found in the Iberian Peninsula [27,28]. Thus, these results can be interpreted as the paternal lineages of El Salvador being admixed from African, Native American, and European sources, with the latter in a greater proportion.

Allele frequencies, a priori statistics, HWE and F_{ST} values for the 94 autosomal identification SNPs in Verogen ForenSeq™ in a population sample of El Salvador are given in Supplementary Table 9. Sample sizes ranged from 392 to 782 chromosomes, with a mean of 754 and a median of 782. For 53 loci, genotypes could be produced for all individuals, and overall 80 SNPs had a fraction of missing genotypes <5%. On the contrary, four loci had >30% missing genotypes: rs1736442 (49.87%), rs2920816 (47.83%), rs7041158 (39.9%), and rs1031825 (31.97%). Eight SNPs failed HWE after Bonferroni correction. Average expected heterozygosity was 0.4406, which is close to the maximum possible value of 0.5. RMP was 3.13×10^{-38}, and the chance of excluding a false father is 1 - 2.28×10^{-8}. When combining the autosomal STRs and SNPs in Verogen ForenSeq™, these a priori statistics take very low values: RMP becomes 4.60×10^{-72}, and the chance of excluding a false father is 1 - 1.06×10^{-20}.

We next computed F_{ST} distances between a set of reference populations: Roma and Catalans from Spain [17] (the former is an ethnic minority, while the latter were sampled as individuals with all four grandparents born in Catalonia), the main ethnic groups in the USA [18], and world populations grouped by continental origins [19]. Table 2 shows the F_{ST} distance matrix computed from the 27 autosomal STRs; in it, Salvadorans appear closest to US Hispanics (F_{ST} = 0.0042), while they are more distant from Catalans, Europeans and European Americans ($F_{ST}\approx$ 0.0200). Intriguingly, the population of El Salvador is slightly closer to European Americans than to Europeans or Catalans from the colonial power, Spain; a possible explanation might be that European Americans are known to carry ~0.18% admixture from Native Americans [29]. Although their distance to Native Americans is larger (F_{ST} = 0.0262), it should be noted that Salvadorans are the population Native Americans are closest to. We applied MDS to this distance matrix, and the result can be seen in Figure 1a. In the plot, El Salvador groups with Hispanics, as well as with East Asians and Asian Americans, even though their F_{ST} values with the latter two populations are relatively high (0.0298-0.0375), and actually higher than the distance between El Salvador and European populations (see above and Table 2). This cluster is relatively close to a European set of populations. The patterns observed with X-STRs are slightly different: while El Salvador is still closest to Hispanics (F_{ST} = 0.0031), it is much closer to Native Americans (F_{ST} = 0.0061) than to populations of European descent (F_{ST} = 0.0242 – 0.0323); as in autosomal STRs, it is closer to European Americans than to populations from Europe. The MDS plot (Figure 1b) reproduces the same general groupings of populations, but Salvadorans, Hispanics and East Asians are closer to Native Americans than they were in the autosomal STR-based plot (Figure 1a).

Discussion

We have typed the 58 STRs and 94 SNPs contained in the Verogen Forenseq™ primer Mix A in 391 samples from El Salvador, and provide allele frequencies for the general population of El Salvador, which will be of invaluable help in the quest for the thousands of people that were disappeared during the El Salvador Civil War (1980-1992), and for identifying human remains of putative migrants to the USA. In particular, the large number of loci contained and the degree of informativity added by sequencing rather than sizing alleles make it particularly adequate when DNA degradation results in a high proportion of missing genotypes. Out of 27

autosomal STRs, we found four (D22S1045, D5S818, PentaD, and PentaE) that failed HWE after Bonferroni correction, in all cases due to a homozygote excess over the expected values under HWE. Technical and/or population cases could have caused this homozygote excess. However, Novroski et al. [18] found 1-3 loci failing HWE in the main ethnic groups in the USA, which are presumed to be large and relatively homogeneous populations. In particular, D5S818 also failed HWE in African Americans, Hispanics and European Americans (although only below the Bonferroni threshold in the latter), as did PentaD in African Americans. This would point to technical causes in heterozygote detection in at least these two STRs. In particular, to date, five sequence variations at the primer binding region of D5S818 have been reported to cause discrepancies in paternity testing since they cause null alleles [30]. And although rarer, null alleles have also been reported for PentaD [31].

As expected given the cultural and ethnic homogeneity of this relatively small country (21,041 Km2, roughly the size of Slovenia, Israel or New Jersey), we found that allele frequencies are not significantly different among the three main regions in El Salvador (Centro, Occidente, and Oriente). In El Salvador, population mobility has always been high, both seasonally (of laborers to the agricultural areas) and permanently (to the capital or abroad) [32]. The war increased migration to the capital, and, while some migrants returned to their home towns, many settled in San Salvador [33]. Still currently, internal migration remains high due to economic or social reasons, such as escaping high-crime areas; the small size of the country implies that it can be covered in a single, countrywide migration network.

Also as expected given that ~90% of the population of El Salvador are Mestizos (the product of historic admixture mostly between Native Americans and Europeans), allele frequencies in El Salvador are closest to those in Hispanics. We should note that we sampled blindly with respect to ethnicity and did not record it; we expect that our sample reflects the average genetic composition of El Salvador. It is noteworthy that, in the case of X-STRs, Salvadorans, Hispanics and other Central Americans [10] are especially close to Native Americans. This is likely the result of the well-known sex-specific admixture patterns in the Americas, where Native American ancestry was contributed mostly by females, while most European migrants were male [34–36]. In the case of El Salvador, Native American mtDNA sequences were found at a ~95% frequency [15], while the Native American Q haplogroup was found at a 31% frequency [12], slightly higher than our 13% STR-based estimate. Since we estimated haplogroup frequencies from Y-STRs rather than from the Y-SNPs that define them, our haplogroup frequencies should be taken with caution[37]. Still, the presence of inferred Q and E1b1a Y chromosomes in our population sample highlights the internal diversity of Y-chromosome haplotypes in El Salvador, which must be taken into account in casework.

It should be noted that we based our population genetics analysis on the standard F_{ST} metric, which does not take into account the mutational distance between alleles. Alternatively, R_{ST} [38] is based on a quantitative variance apportionment of allele size, and it reflects better the evolutionary history of size-based alleles. One would expect that, besides the effects of genetic drift, founder effects and gene flow that can be expected in an admixed population such that of El Salvador, and that can be captured by F_{ST}, R_{ST} would add the mutational history, that is, the fact that presumably two populations with smaller allele size differences are more closely related to each other than populations with larger allele size differences. However, massive

parallel sequencing of STR alleles has uncovered different layers of complexity in STR structure, comprising different repeat arrays, single nucleotide mutations or repeat conversions, which imply that isometric alleles cannot be considered as single evolutionary entities. Thus, it would be desirable to implement some ad hoc metric of evolutionary distance among RSB alleles, which would probably need to be locus-specific. Instead, as a basic phenetic distance, we have used F_{ST}, which, as detailed in the paragraph above, has allowed us to retrieve the expected population genetic patterns of Salvadorans, and which may imply that the time scale of differentiation by mutation is deeper than that caused by drift and admixture in mixed American populations.

ACKNOWLEDGEMENTS

This article is dedicated to the memory of the late Cristián Orrego Benavente, who had this initiative, and for his general contribution to the defense of human rights in El Salvador.

We would like to particularly thank all the volunteers participating in this study. We are particularly grateful to (now deceased) for this initiative and for. This work was supported by the Spanish Ministry of Economy and Competitiveness and Agencia Estatal de Investigación (grant numbers CGL2016-75389-P (MINEICO/FEDER, UE), PID2019-106485GB-I00/AEI/10.13039/501100011033 (MINEICO), and “Unidad María de Maeztu” (MDM-2014-0370) to FCal; Agència de Gestió d’Ajuts Universitaris i de la Recerca (Generalitat de Catalunya, grant 2017SGR00702); Agència Catalana de Cooperació al Desenvolupament (ACCD004/17/00019 and ACCD016/18/00031); Fundación Panamericana para el Desarrollo (PADF, No. PRDHD-RFA-R-2017-009). We thank also the Ministry of Health of El Salvador, which, in 2018, allowed us to take samples at their facilities.

REFERENCES

[1] DIGESTYC, Encuesta de Hogares de Hogares Múltiples, Ministerio de Economía, San Salvador, 2020.

[2] J. Lemus, Sociolinguistic Atlas of Indigenous Peoples in Latin America, UNICEF and FUNPROEIB, Cochabamba, Bolivia, 2009.

[3] Centre for the Autonomy and Development of Indigenous Peoples updated by IFAD, Country technical note on indigenous peoples' issues, Republic of El Salvador, San Salvador, 2017.

[4] P. Muñoz, E.L. Pinto de Erazo, C. Baeza, E. Arroyo-Pardo, A.M. López-Parra, Genetic polymorphism of 15 STR loci in El Salvador, Int. J. Legal Med. 129 (2015) 991–993. https://doi.org/10.1007/s00414-015-1148-8.

[5] J.C. Monterrosa, J. Morales, I. Yurrebaso, O. García, Population genetic data for 16 STR loci (PowerPlex ESX-17 kit) in El Salvador, Forensic Sci. Int. Genet. 6 (2012). https://doi.org/10.1016/j.fsigen.2011.12.004.

[6] J. Lovo-Gómez, A. Salas, Á. Carracedo, Microsatellite autosomal genotyping data in four indigenous populations from El Salvador, Forensic Sci. Int. 170 (2007) 86–91. https://doi.org/10.1016/j.forsciint.2006.05.031.

[7] J.C. Monterrosa, J.A. Morales, O. García, Genetic variation for 15 short tandem repeat loci in an El Salvadoran (Central America) population, J. Forensic Sci. 51 (2006) 451–452. https://doi.org/10.1111/j.1556-4029.2006.00097.x.

[8] B. Martínez-Jarreta, P. Vásquez, E. Abecia, M. Garde, I. de Blás, B. Budowle, Autosomic STR Loci (HUMTPOX, HUMTH01, HUMVWA, D18S535, D1S1656 and D12S391) in San Salvador (El Salvador, Central America), J. Forensic Sci. 49 (2004) 1–2. https://doi.org/10.1520/jfs2003395.

[9] J.A. Morales, J.C. Monterrosa, J.C. Alvarez, C. Entrala, J.A. Lorente, M. Lorente, B. Budowle, E. Villanueva, Population Data on Nine STR Loci in an El Salvadoran (Central American) Sample Population, J. Forensic Sci. 47 (2002) 15461J. https://doi.org/10.1520/jfs15461j.

[10] M. Baeta, E. Prieto-Fernández, C. Núñez, T. Kleinbielen, P. Villaescusa, L. Palencia-Madrid, O. Alvarez-Gila, B. Martínez-Jarreta, M.M. de Pancorbo, Study of 17 X-STRs in Native American and Mestizo populations of Central America for forensic and population purposes, Int. J. Legal Med. (2021). https://doi.org/10.1007/s00414-021-02536-9.

[11] J.C. Monterrosa, J.A. Morales, I. Yurrebaso, L. Gusmão, O. García, Population data for 12 Y-chromosome STR loci in a sample from El Salvador, Leg. Med. 12 (2010) 46–51. https://doi.org/10.1016/j.legalmed.2009.10.003.

[12] J. Lovo-Gómez, A. Blanco-Verea, M. V. Lareu, M. Brión, A. Carracedo, The genetic male legacy from El Salvador, Forensic Sci. Int. 171 (2007) 198–203. https://doi.org/10.1016/j.forsciint.2006.07.005.

[13] B. Martínez-Jarreta, P. Vásquez, E. Abecia, B. Budowle, A. Luna, F. Peiró, Characterization of 17 Y-STR Loci in a Population from El Salvador (San Salvador, Central America) and Their Potential for DNA Profiling, J. Forensic Sci. 50 (2005) 1–4.

https://doi.org/10.1520/jfs2005173.

[14] J. Saul, M. Fondevila, A. Salas, M. Brión, M.V. Lareu, Á. Carracedo, Y-chromosome STR-haplotype typing in El Salvador, Forensic Sci. Int. 142 (2004) 45–49. https://doi.org/10.1016/j.forsciint.2004.02.004.

[15] A. Salas, J. Lovo-Gómez, V. Álvarez-Iglesias, M. Cerezo, M.V. Lareu, V. Macaulay, M.B. Richards, Á. Carracedo, Mitochondrial echoes of first settlement and genetic continuity in El Salvador, PLoS One. 4 (2009) e6882. https://doi.org/10.1371/journal.pone.0006882.

[16] F. Calafell, R. Anglada, N. Bonet, M. González-Ruiz, G. Prats-Muñoz, R. Rasal, C. Lalueza-Fox, J. Bertranpetit, A. Malgosa, F. Casals, An assessment of a massively parallel sequencing approach for the identification of individuals from mass graves of the Spanish Civil War (1936–1939), Electrophoresis. 37 (2016). https://doi.org/10.1002/elps.201600180.

[17] 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). https://doi.org/10.1016/j.fsigen.2017.06.006.

[18] 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. https://doi.org/10.1016/j.fsigen.2016.09.007.

[19] C. Phillips, L. Devesse, D. Ballard, L. van Weert, M. de la Puente, S. Melis, V. Álvarez Iglesias, A. Freire-Aradas, N. Oldroyd, C. Holt, D. Syndercombe Court, Á. Carracedo, M.V. Lareu, Global patterns of STR sequence variation: Sequencing the CEPH human genome diversity panel for 58 forensic STRs using the Illumina ForenSeq DNA Signature Prep Kit, Electrophoresis. 39 (2018) 2708–2724. https://doi.org/10.1002/elps.201800117.

[20] 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 FGx™ forensic genomics system, Forensic Sci. Int. Genet. 24 (2016) 18–23. https://doi.org/10.1016/j.fsigen.2016.05.008.

[21] 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 (2010) 564–567.

[22] R.A. Fisher, Standard calculations for evaluating a blood system, Heredity (Edinb). 5 (1951) 95–102.

[23] P.E. Smouse, R. Chakraborty, The use of restriction fragment length polymorphisms in paternity analysis, Am. J. Hum. Genet. 38 (1986) 918–939. https://pubmed.ncbi.nlm.nih.gov/3014872/ (accessed November 5, 2021).

[24] M. van Oven, A. Van Geystelen, M. Kayser, R. Decorte, M.H.D. Larmuseau, Seeing the wood for the trees: a minimal reference phylogeny for the human Y chromosome., Hum. Mutat. 35 (2014) 187–91. https://doi.org/10.1002/humu.22468.

[25] 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. https://doi.org/10.1016/J.FSIGEN.2016.06.008.

[26] L. Gusmão, J.M. Butler, A. Linacre, W. Parson, W. Parson, P.M. Schneider, A. Carracedo, Revised guidelines for the publication of genetic population data, Forensic Sci. Int. Genet. 30 (2017) 160–163. https://doi.org/10.1016/J.FSIGEN.2017.06.007.

[27] S.M.M. Adams, E. Bosch, P.L.L. Balaresque, S.J.J. Ballereau, A.C.C. Lee, E. Arroyo, A.M. López-Parra, M. Aler, M.S.G.S. Grifo, M. Brion, A. Carracedo, J. Lavinha, B. Martínez-Jarreta, L. Quintana-Murci, A. Picornell, M. Ramon, K. Skorecki, D.M.M. Behar, F. Calafell, M.A.A. Jobling, A.M. Lopez-Parra, M. Aler, M.S.G.S. Grifo, M. Brion, A. Carracedo, J. Lavinha, B. Martinez-Jarreta, L. Quintana-Murci, A. Picornell, M. Ramon, K. Skorecki, D.M.M. Behar, F. Calafell, M.A.A. Jobling, The genetic legacy of religious diversity and intolerance: paternal lineages of Christians, Jews, and Muslims in the Iberian Peninsula, Am. J. Hum. Genet. 83 (2008) 725–736. https://doi.org/10.1016/j.ajhg.2008.11.007.

[28] N. Solé-Morata, J. Bertranpetit, D. Comas, F. Calafell, Y-chromosome diversity in Catalan surname samples: insights into surname origin and frequency., Eur. J. Hum. Genet. 23 (2015) 1549–57. https://doi.org/10.1038/ejhg.2015.14.

[29] K. Bryc, E.Y. Durand, J.M. Macpherson, D. Reich, J.L. Mountain, The genetic ancestry of African Americans, Latinos, and European Americans across the United States, Am. J. Hum. Genet. 96 (2015) 37–53. https://doi.org/10.1016/J.AJHG.2014.11.010.

[30] C. Shao, Y. Yao, X. Pan, M. Wu, B. Zhang, H. Xu, J. Xie, K. Sun, Variants in linkage status at D5S818 detected by multiple STR kits comparison and Sanger sequencing, Mol. Genet. Genomic Med. 9 (2021). https://doi.org/10.1002/MGG3.1765.

[31] K.J. Van Der Gaag, R.H. De Leeuw, J. Hoogenboom, J. Patel, D.R. Storts, J.F.J. Laros, P. De Knijff, Massively parallel sequencing of short tandem repeats—Population data and mixture analysis results for the PowerSeq™ system, Forensic Sci. Int. Genet. 24 (2016) 86–96. https://doi.org/10.1016/J.FSIGEN.2016.05.016.

[32] S. Montes, Displaced persons and Salvadoran refugees, Int. Relations. Natl. Univ. Costa Rica. 13 (1985) 11–21.

[33] J.D. Morán Mendoza, Guerra y migración interna en El Salvador, in: P.C. de Población (Ed.), Semin. Int. Población Del Istmo Al Final Del Milen., San José, Costa Rica, 1999: pp. 307–333.

[34] I. Mendizabal, K. Sandoval, G. Berniell-Lee, F. Calafell, A. Salas, A. Martinez-Fuentes, D. Comas, A. Martínez-Fuentes, D. Comas, Genetic origin, admixture, and asymmetry in maternal and paternal human lineages in Cuba, BMC Evol Biol. 8 (2008) 213. https://doi.org/10.1186/1471-2148-8-213.

[35] L. Ongaro, M.O. Scliar, R. Flores, A. Raveane, D. Marnetto, S. Sarno, G.A. Gnecchi-Ruscone, M.E. Alarcón-Riquelme, E. Patin, P. Wangkumhang, G. Hellenthal, M. Gonzalez-Santos, R.J. King, A. Kouvatsi, O. Balanovsky, E. Balanovska, L. Atramentova, S. Turdikulova, S. Mastana, D. Marjanovic, L. Mulahasanovic, A. Leskovac, M.F. Lima-Costa, A.C. Pereira, M.L. Barreto, B.L. Horta, N. Mabunda, C.A. May, A. Moreno-Estrada, A. Achilli, A. Olivieri, O. Semino, K. Tambets, T. Kivisild, D. Luiselli, A. Torroni, C. Capelli,

E. Tarazona-Santos, M. Metspalu, L. Pagani, F. Montinaro, The Genomic Impact of European Colonization of the Americas, Curr. Biol. 29 (2019) 3974-3986.e4. https://doi.org/10.1016/j.cub.2019.09.076.

[36] A. Moreno-Estrada, S. Gravel, F. Zakharia, J.L. McCauley, J.K. Byrnes, C.R. Gignoux, P.A. Ortiz-Tello, R.J. Martínez, D.J. Hedges, R.W. Morris, C. Eng, K. Sandoval, S. Acevedo-Acevedo, P.J. Norman, Z. Layrisse, P. Parham, J.C. Martínez-Cruzado, E.G. Burchard, M.L. Cuccaro, E.R. Martin, C.D. Bustamante, Reconstructing the Population Genetic History of the Caribbean, PLoS Genet. 9 (2013). https://doi.org/10.1371/journal.pgen.1003925.

[37] J. Jannuzzi, J. Ribeiro, C. Alho, G. de Oliveira Lázaro e Arão, R. Cicarelli, H. Simões Dutra Corrêa, S. Ferreira, C. Fridman, V. Gomes, S. Loiola, M.F. da Mota, Â. Ribeiro-dos-Santos, C.A. de Souza, R.S. de Sousa Azulay, E.F. Carvalho, L. Gusmão, Male lineages in Brazilian populations and performance of haplogroup prediction tools, Forensic Sci. Int. Genet. 44 (2020). https://doi.org/10.1016/J.FSIGEN.2019.102163.

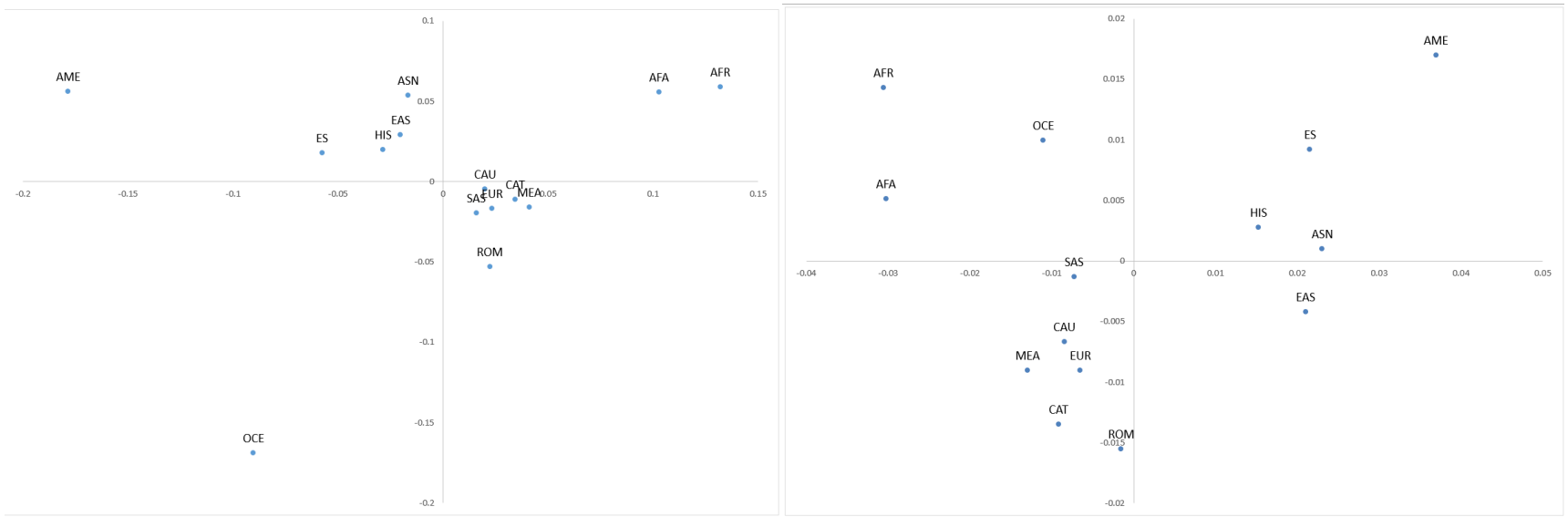
[38] M. Slatkin, A measure of population subdivision based on microsatellite allele frequencies, 139 (1995) 457–462.

Haplogroup	N	Relative freq. (%)
E1b1a	4	5.3
E1b1b-M35	1	1.3
E1b1b-M78	9	11.8
E1b1b-V22	1	1.3
E1b1b-M123	2	2.6
E1b1b-M81	2	2.6
G2a2b-M406	1	1.3
I1-M253	1	1.3
I2-M26	3	3.9
I2-L460	1	1.3
I2-L596	1	1.3
J1-P58	1	1.3
J2a-L26	1	1.3
J2a-M319	1	1.3
J2a-L25	1	1.3
J2a-L581	1	1.3
Q1a2-M346	10	13.2
R1a	2	2.6
R1b	30	39.5
T	2	2.6
Unknown	1	1.3

Table 1. Relative haplogroup frequencies (*Relative freq.*) predicted from Y-STRs using the Bayesian predictor nevgen [www.nevgen.org] for the 76 complete Y-STR haplotypes in samples from El Salvador. In the case labelled “Unknown”, no predicted haplotype reached a posterior probability > 80%

	ES	CAT	ROM	AFA	AFR	ASN	EAS	CAU	EUR	MEA	HIS	AME	OCE	SAS
ES	0	0.0323	0.0278	0.0389	0.0388	0.0109	0.0121	0.0242	0.0265	0.0310	0.0031	0.0061	0.0315	0.0230
CAT	0.0212	0	0.0102	0.0234	0.0319	0.0322	0.0271	0.0027	0.0022	0.0032	0.0202	0.0500	0.0188	0.0075
ROM	0.0250	0.0146	0	0.0271	0.0327	0.0279	0.0213	0.0090	0.0094	0.0136	0.0173	0.0471	0.0126	0.0111
AFA	0.0338	0.0253	0.0313	0	0.0044	0.0449	0.0435	0.0175	0.0213	0.0195	0.0337	0.0605	0.0191	0.0176
AFR	0.0509	0.0399	0.0455	0.0083	0	0.0454	0.0447	0.0197	0.0260	0.0255	0.0383	0.0619	0.0187	0.0213
ASN	0.0298	0.0268	0.0288	0.0287	0.0423	0	0.0007	0.0251	0.0261	0.0313	0.0100	0.0147	0.0239	0.0215
EAS	0.0375	0.0314	0.0351	0.0369	0.0388	0.0090	0	0.0222	0.0197	0.0260	0.0081	0.0200	0.0237	0.0196
CAU	0.0197	0.0012	0.0141	0.0242	0.0372	0.0231	0.0277	0	0.0024	0.0060	0.0150	0.0423	0.0122	0.0032
EUR	0.0219	0.0021	0.0151	0.0264	0.0355	0.0275	0.0256	0.0029	0	0.0028	0.0163	0.0438	0.0128	0.0059
MEA	0.0267	0.0079	0.0142	0.0208	0.0279	0.0244	0.0244	0.0085	0.0047	0	0.0210	0.0489	0.0175	0.0050
HIS	0.0042	0.0156	0.0200	0.0274	0.0433	0.0223	0.0320	0.0119	0.0155	0.0198	0	0.0121	0.0293	0.0163
AME	0.0262	0.0644	0.0600	0.0710	0.0759	0.0613	0.0529	0.0580	0.0573	0.0618	0.0317	0	0.0489	0.0377
OCE	0.0499	0.0639	0.0611	0.0670	0.0678	0.0639	0.0570	0.0590	0.0523	0.0562	0.0552	0.0675	0	0.0095
SAS	0.0248	0.0122	0.0145	0.0244	0.0311	0.0189	0.0156	0.0107	0.0072	0.0076	0.0197	0.0544	0.0459	0

Table 2. F_{ST} distance matrices based on 27 autosomal STRs (below diagonal) or 7 X-STRs (above diagonal). ES, El Salvador (present data). CAT, Catalans; ROM, Spanish Roma [17]. AFA, African Americans; ASN, Asian Americans; CAU, *Caucasians*; HIS, Hispanics [18]. AFR, Africans; EAS, East Asians; EUR, Europeans; MEA, Middle Easterners and North Africans; AME, Native Americans; OCE, Oceanians; SAS, South Asians [19].



a) b)

Figure 1. Multidimensional scaling (MDS) plots based on F_{ST} distances from a) 27 autosomal STRs and b) 7 X-STRs. as in Table 2. Stress values are a) 12.1%, b) 7.0%. In MDS, stress values measure the degree of concordance between the original distance (F_{ST} in this case) and the distances between points as they appear in the plot; stress values closer to 0 signify a better concordance. Population abbreviations: ES, El Salvador; CAT, Catalans; ROM, Spanish Roma; AFA, African Americans; ASN, Asian Americans; CAU, *Caucasians*; HIS, Hispanics; AFR, Africans; EAS, East Asians; EUR, Europeans; MEA, Middle Easterners and North Africans; AME, Native Americans; OCE, Oceanians; SAS, South Asians. Note the position of ES closer to Hispanics, and in b), to Native Americans as well.

Supplementary Table 1. Repeat sequence based allele structure and nomenclature for the STRs in the Verogen Forenseq™ Primer Mix A. The number and sequence of the repetitive units in each allele are indicated. LB allele: length-based allele names; RSB allele: repeat-sequence based allele names; sequence: repeat-region sequence region, as provided by the Verogen Forenseq™ UAS.

Supplementary Table 2. Length-based absolute (N) and relative (rel) allele frequencies, observed heterozygosity (Obs. Het.), expected heterozygosity (Exp. Het.), HWE p-value (p HWE), power of discrimination (POD), chance of exclusion in a paternity trio (CE), F_{ST} and the p-value for F_{ST} (p Fst) in 27 autosomal STRs in a population sample of 391 individuals from El Salvador. The number of individuals for which we could obtain a genotype is indicated next to each locus name.

Supplementary Table 3. Repeat sequence-based absolute (N) and relative (rel) allele frequencies, observed heterozygosity (Obs. Het.), expected heterozygosity (Exp. Het.), HWE p-value (p HWE), power of discrimination (POD), chance of exclusion in a paternity trio (CE), F_{ST} and the p-value for F_{ST} (p Fst) in 27 autosomal STRs in a population sample of 391 individuals from El Salvador.. See Supplementary Table 1 for allele sequences. The number of individuals for which we could obtain a genotype is indicated next to each locus name.

Supplementary Table 4. Length-based allele absolute (N) and relative (rel) frequencies, observed heterozygosity (Obs. Het.), expected heterozygosity (Exp. Het.) (both in females), HWE p-value (p HWE), F_{ST} and the p-value for F_{ST} (p Fst) in 7 X-STRs, in a population sample of 534 X chromosomes from El Salvador The number of individuals for which we could obtain a genotype is indicated next to each locus name.

Supplementary Table 5. Repeat sequence-based absolute (N) and relative (rel) allele frequencies, observed heterozygosity (Obs. Het.), expected heterozygosity (Exp. Het.) (both in females), HWE p-value (p HWE), F_{ST} and the p-value for F_{ST} (p Fst) in 7 X-STRs, in a population sample of 534 X chromosomes from El Salvador. See Supplementary Table 1 for allele sequences. The number of individuals for which we could obtain a genotype is indicated next to each locus name.

Supplementary Table 6. Length-based haplotype frequencies for the DXS10135-DXS8378, DXS7132-DXS10074, and DXS10103-HPRTB pairs. Haplotype frequencies were estimated by direct counting in males and informative (i.e, heterozygote in at most one locus within a particular pair) females. Exp. Het.: expected heterozygosity.

Supplementary Table 7. Repeat sequence-based haplotype frequencies for the DXS10135-DXS8378, DXS7132-DXS10074, and DXS10103-HPRTB pairs. See Supplementary Table 1 for allele sequences. Haplotype frequencies were estimated by direct counting in males and informative (i.e, heterozygote in at most one locus within a particular pair) females. Exp. Het.: expected heterozygosity.

Supplementary Table 8. Repeat sequence-based Y-STR haplotype frequencies, for the 76 Y chromosomes from El Salvador for which we could genotype the complete set of Y-STRs in the

Verogen Forenseq™ Primer Mix A. See Supplementary Table 1 for allele sequences. Abs. freq.: absolute frequency; Rel. freq.: relative frequency.

Supplementary Table 9. Allele frequencies, observed heterozygosity (Obs. Het.), expected heterozygosity (Exp. Het.), HWE p-value (p HWE), power of discrimination (PD), chance of exclusion in a paternity trio (CE), F_{ST} and the p-value for F_{ST} (p Fst) in the 94 SNPs in the Verogen Forenseq™ Primer Mix A in a population sample of 391 individuals from El Salvador. N: number of individuals for which we could obtain a genotype for each locus.