ORIGINAL ARTICLE

Forensic efficiency of microsatellites and single nucleotide polymorphisms on the X chromosome

María T. Zarrabeitia · Verónica Mijares ·
José A. Riancho

Received: 10 November 2006 / Accepted: 20 March 2007 / Published online: 14 April 2007

Abstract Polymorphisms located on the X chromosome are being increasingly used in forensic studies. However, they have not been studied as extensively as the autosomal and Y-linked polymorphisms. Therefore, we planned to study the population genetics of two sets of X-linked markers, including short-tandem repeats (STRs) and single nucleotide polymorphisms (SNPs), and particularly analyze the differences at the microgeographical level. Six X-linked STRs (DXS 9895, DXS 7132, DXS 9898, DXS 6789, GATA 172D05, and DXS 7130) and ten SNPs (rs1229078, rs1544545, rs4442270, rs1874111, rs5968332, rs1166756, rs12849634, rs5932595, rs203648, and rs611711) were studied in two population samples from Cantabria, northern Spain, a mixed coastal population and a relatively isolated small population in the Pas valley. There were statistically significant differences in allelic frequencies of the six STRs studied between both populations. On the other hand, only one out of ten SNPs studied showed between-population differences. Overall Fst values were 0.4–2.9% and 1.8–2.2% for the STRs and the SNPs, respectively. The overall power of discrimination for female samples was higher than 99.99% for both groups of markers. Therefore, these sets of X-linked STRs and SNPs seem to be potentially useful in forensic genetics, but care should be taken when interpreting results from cases that originate from small and relatively isolated populations.

Keywords Short tandem repeats · Single nucleotide polymorphisms · X chromosome · Population data

Electronic supplementary material The online version of this article (doi:10.1007/s00414-007-0169-3) contains supplementary material, which is available to authorized users.

M. T. Zarrabeitia (✉) · V. Mijares
Unit of Legal Medicine, University of Cantabria,
Av. Herrera Oria sn,
39011, Santander, Spain
e-mail: zarrabet@unican.es

J. A. Riancho
Department of Internal Medicine, Hospital U.M. Valdecilla,
University of Cantabria,
Santander, Spain

Introduction

To interpret the results of DNA analysis in a forensic case, they need to be compared with a pertinent reference population. Fortunately, the allelic frequency distributions of most autosomal short tandem repeats (STR) do not show great variations within large ethnic groups (i.e., Caucasians) [6]. However, that may be not the case for loci located on sex chromosomes and particularly for Y-linked markers. The smaller effective population and the lack of recombination make them more prone to show different frequency distributions related to population structure. Indeed, we have shown evidence for population differentiation at the microgeographical level in small rural areas when Y-linked STRs are studied [32].

STRs of the X chromosome are being increasingly studied as a useful tool in forensic medicine [5, 10, 26, 27, 30, 31]. The effective population size of X chromosomes is intermediate between those of autosome pairs and the Y chromosome (there is one Y chromosome and three X chromosomes for every four autosomal chromosomes). We and others have recently shown that the allelic frequencies of X-linked loci do not show great variability on a nationwide basis thus allowing general databases to be used as references [5, 29]. However, there are only scarce empirical data about the population genetics of X-linked STRs within the regional level. This is important for several reasons. First, that information is needed to determine the suitability of general databases for reference in cases of relatively isolated communities. Second, although even relatively close loci do not seem to show a significant degree of linkage disequilibrium in the general population, it is unclear what the situation is in small and relatively isolated populations [13, 14, 22]. This is a critical issue in forensic medicine, as it dictates whether the statistical calculations are to be made combining the results derived from several individual loci, each one being treated as independent, or if it must be done at the "haplotype" level.

Therefore, we planned this study to analyze the population genetics, particularly the allele frequency distribution and the linkage disequilibrium, of X-linked STRs and single nucleotide polymorphisms (SNPs) in two population groups representing a well-mixed population and a small and relatively isolated one.

Materials and methods

Population

We studied 377 individuals from Cantabria, 227 from the coastal area (142 men and 85 women), and 150 from the Pas valley (95 men and 55 women). Cantabria is a 5,000-km^2 region in northern Spain, situated between the sea and the Cantabrian Mountains, with a total population of 530,000. Most of the population lives in the well-connected coastal area, which has a mixed and relatively mobile population of about 400,000. On the other hand, the southern part of Cantabria is a mountainous area with several valleys that traditionally had difficult communication, and the inhabitants had less opportunity for social interaction with other areas. That has been particularly marked for the population of the Pas valley, a mostly cattle-raising community that traditionally had a short-distance seminomad life, moving with cattle from the valley floor in winter to the nearby hills in spring and summer [11, 15]. Population data for STRs in a partially overlapping group in the coastal region have been previously published [29].

DNA analysis

Genomic DNA was extracted from peripheral blood by a commercial method following the manufacturer's instructions (Qiamp DNA blood mini kit; Qiagen, Hilden, Germany) and quantified by light absorbance.

STRs

Six tetranucleotide X-STRs (DXS 9895, DXS 7132, DXS 9898, DXS 6789, GATA 172D05, and DXS 7130; Fig. 1) were studied by polymerase chain reaction (PCR) and capillary electrophoresis as previously reported [5, 29]. We also typed the DNA sample NA9947 (available in the Powerplex kit from Promega, Madison, WI) as a control. The results were in agreement with those reported by Szibor et al. [24].

SNPs

A set of ten SNPs was chosen (Fig. 1). To select individual SNPs, we explored genome databases looking for SNPs with balanced allelic frequencies (to maximize discriminating power) distributed through the whole X chromosome, excluding the pseudoautosomal regions. Genotyping was done with Taqman assays using 10–20 ng DNA, primers, and allele-specific probes designed and supplied by Applied Biosystems (Foster City, CA) and a universal PCR master mix in a total reaction volume of 5 μl with the cycling conditions recommended by the manufacturer. The following assays were used: rs1229078 (C_8817140_10), rs1544545 (C_8821450_10), rs4442270 (C_111236_10), rs1874111 (C_12118788_10), rs5968332 (C_2465330_10), rs1166756 (C_11382383_10), rs12849634 (C_11174579_10), rs5932595 (C_2465719_10), rs203648 (C_3085402_10), and rs611711(C_1089522_1).

Statistics

Population differences in allelic frequencies were tested by a Monte Carlo extension of the Fisher exact test using the SPSS software (SPSS, Chicago, IL) and 10,000 permutations. Hardy-Weinberg equilibrium (HWE), linkage disequilibrium, and population stratification (determined by the coancestry coefficient Fst, representing allele identity by descent between individuals within subpopulations relative to the total) were estimated with GDA software (Lewis and Zaykin; Genetic Data Analysis, v.1.1. http://lewis.eeb.uconn.edu/lewishome/software.html), which implements the formulae published by Weir [28], and with the Arlequin v3.0 (Excoffier et al. Arlequin v 3.0—An integrated software package for population genetics data analysis; http://cmpg.unibe.ch/software/ arlequin3).

The false discovery rate strategy was used to assess the statistical significance of p values after correction for multiple comparisons [3, 4].

Locus diversity was calculated as $\left(1-\sum \mathrm{Pi}^2\right)\times n/(n-1)$, where Pi is the allele frequency. The forensic efficiency was calculated with the formulae given by Desmarais et al. [8].

Results

STRs

There were no sex-related differences in allele frequencies. Hence, combined data from males and females are shown.

Table 1 Forensic efficiency of X-linked STRs

STR	PDf		PDm		PEtrio		PEmo	
	Coast	Pas	Coast	Pas	Coast	Pas	Coast	Pas
DXS9895	0.895	0.862	0.751	0.710	0.708	0.656	0.569	0.511
DXS7132	0.891	0.906	0.744	0.759	0.701	0.723	0.562	0.587
DXS9898	0.914	0.883	0.778	0.735	0.742	0.688	0.609	0.547
DXS6789	0.903	0.878	0.754	0.717	0.717	0.675	0.582	0.535
G172D05	0.932	0.935	0.804	0.808	0.774	0.780	0.648	0.656
DXS7130	0.881	0.864	0.718	0.712	0.678	0.659	0.538	0.515
Combined (%)	99.9999	99.9998	99.9807	99.9710	99.9536	99.9273	99.4991	99.2909

PDf Power of discrimination in female cases; *PDm* power of discrimination in male cases; *PEtrio* power of exclusion of paternity in trio cases; *PEmo* power of exclusion of paternity in motherless cases

There were statistically significant differences in allele frequencies between the coastal region and the Pas valley for all six STRs studied. Fst values tended to be somewhat higher in females than in males, with overall values over the six loci of 0.4 and 2.9%, respectively (see Tables 1E and 2E in the electronic supplementary material for this article).

Allele frequencies in women did not depart from the HWE. Likewise, we did not find statistical evidence for linkage disequilibrium between the different loci. Although some of the comparisons between different loci pairs were associated with unadjusted *p* values less than 0.05, all were above the significant threshold after adjustment for multiple comparisons by the false discovery rate procedure.

The forensic efficiency of the STRs is shown in Table 1. It was somewhat lower in the Pas valley than in the coastal population, but the difference was small. The combined power of discrimination was above 99.97% in both cases, for female as well as for male samples. Likewise, the combined power of exclusion of paternity in standard trio cases was 99.92–99.95%, which decreased to 99.29–99.49% in the absence of maternal data.

SNPs

There were no differences in allelic frequencies between the male and female subgroups. On the other hand, one of the ten SNPs studied (rs1166756) showed significant between-population differences in allelic frequencies, with an unadjusted *p* value <0.00001 well below the threshold of 0.005 needed to maintain a global false discovery error rate of 0.05 (Table 3E in the electronic supplementary material). The overall Fst value was 2.1% in females and 1.8% in males.

Female allelic frequencies did not depart from the HWE. Only one between-loci pairwise comparison was associated with significant *p* values suggesting the existence of disequilibrium (rs5932595/rs5968332 in the Pas female group; p=0.0009), despite the fact that both SNPs are about 40 Mb apart and therefore are not physically linked.

Forensic efficiency data are shown in Table 2. As expected from the smaller gene diversity, SNPs had a lower forensic efficiency than STRs. The combined power of discrimination was higher than 99.8% for female samples

Table 2 Forensic efficiency of X-linked SNPs

SNP	PDf		PDm		PEtrio		PEmo	
	Coast	Pas	Coast	Pas	Coast	Pas	Coast	Pas
rs1229078	0.617	0.619	0.485	0.488	0.367	0.369	0.242	0.244
rs1544545	0.603	0.623	0.460	0.497	0.354	0.373	0.230	0.248
rs4442270	0.614	0.625	0.480	0.5000	0.365	0.375	0.240	0.250
rs1874111	0.625	0.615	0.500	0.481	0.375	0.365	0.250	0.241
rs5968332	0.601	0.613	0.613	0.477	0.477	0.363	0.363	0.239
rs1166756	0.623	0.563	0.496	0.404	0.373	0.322	0.248	0.202
rs12849634	0.625	0.698	0.500	0.549	0.375	0.450	0.250	0.314
rs5932595	0.618	0.604	0.486	0.462	0.368	0.355	0.243	0.231
rs203648	0.624	0.600	0.498	0.456	0.374	0.352	0.249	0.228
rs611711	0.576	0.602	0.420	0.459	0.332	0.354	0.210	0.230
Combined (%)	99.9924	99.9933	99.8519	99.8514	99.4405	98.9952	93.5023	93.8330

Abbreviations as in Table 1

as well as for male samples. Likewise, the combined power of exclusion of paternity was 98.399–99.44% in trio cases and decreased to 93.50–93.83% when maternal data were not available.

Discussion

The statistical interpretation of the results of STR analysis relies on the comparison of the case profile with an adequate reference database. It is widely accepted that general databases are usually adequate when studying autosomal STRs, as they show little variation among different population groups of the same ethnic background (e.g., Caucasian) [6, 12].

However, sex chromosomes may be more prone to show population differences. Y chromosomes do not recombine with the X chromosome (except in the short telomeric regions), and their number in the population is only one quarter of the number of each autosomal chromosome. Thus, Y-linked markers are more likely than the autosomal ones to show population structure-related differences [32].

X chromosomes have characteristics somewhere between the autosomes and the Y chromosome. It is thought that the two sex chromosomes, X and Y, diverged from a single autosome about 300 million years ago. Since then, every existing X chromosome has spent two thirds of its history in females. As a consequence, given the lower mutation rates in females than in males, mutations in the X chromosome would tend to occur somewhat less frequently than in the autosomes and the Y chromosome, which explains its lower genetic diversity [19, 20]. On the other hand, the effective population of the X chromosomes is three quarters of the autosomal chromosomes, and they can recombine, but only in female meioses not in male meioses. Therefore, theoretically, the value of X chromosomes to show differentiation between populations would be expected to be intermediate between those of autosomal and Y chromosomes. In fact, that seems to be the case according to the results of several studies [7, 20, 21] and the comparison of data in the present paper and our previous studies of autosomal and Y-linked markers in these populations [32].

In this study, we found significant differences in STR allelic frequencies between the coastal (mixed) population and the relatively isolated population of the Pas valley. Such results contrast with the overall homogeneity of allelic frequencies observed in a nationwide study [29] and suggest that data should be interpreted cautiously when cases from small relatively isolated groups are studied. We found statistically significant between-population differences in all the STRs studied, but such differences were found only in one out of ten SNPs. Although this could be the consequence of the larger allelic variability of the STRs, it may also reflect the higher mutation rate of STRs, which make them more likely to be influenced by the recent history of the populations. On the other hand, the allelic frequency distribution in our coastal population was similar to that reported in other European populations [9, 16, 17], but it was rather different from the allelic frequencies in Oriental studies [2].

Autosomal STRs are the markers of choice for most case studies in the field of forensic genetics. However, X-linked markers are particularly useful in some situations, as reviewed by Szibor et al. [26]. These include paternity deficiency cases when the alleged father is absent and there are two possible daughters, paternity cases involving close blood relatives, maternity cases, and some identification cases. The discrimination power of individual SNPs is much lower than that of STRs. Hence, STRs are usually preferred for forensic analysis. However, SNPs may be advantageous in some situations, given the lower mutation rate and an easier amplification in degraded samples [23]. The amplification of shorter fragments containing repeat polymorphisms (“mini STRs”) can also be of particular interest in samples with low-quality DNA [1]. The procedure to type SNPs used in this study is convenient when there is a relatively large availability of DNA. If only limited amounts are available, it may be necessary to use other techniques involving multiplex reactions or to perform a whole genome amplification of the sample previously. Nevertheless, in our experience, the Taqman technique can be easily scaled down to use 2 ng DNA (Zarrabeitia et al., unpublished observations).

In this study, we present sets of X-linked STRs and SNPs that give reasonable discrimination and paternity exclusion power. They do not show significant linkage disequilibrium in these populations, and therefore, calculations could combine single markers treating them as independent. Nevertheless, this sort of analysis should be interpreted cautiously when several markers on the same chromosome are combined, particularly when forensic cases originate from isolated populations, which are more prone to show between-loci linkage disequilibrium [13, 14]. In case of doubt, the more conservative “haplotypic approach” can be used. On the other hand, studying several markers strongly linked within a haplotype block may be useful in certain kinship cases [18, 25].

References

1. Asamura H, Sakai H, Kobayashi K, Ota M, Fukushima H (2006) MiniX-STR multiplex system population study in Japan and application to degraded DNA analysis. Int J Legal Med 120:174–181
2. Asamura H, Sakai H, Ota M, Fukushima H (2006) Japanese population data for eight X-STR loci using two new quadruplex systems. Int J Legal Med 120:303–309

3. Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I (2001) Controlling the false discovery rate in behavior genetics research. Behav Brain Res 125:279–284
4. Benjamini Y, Yekutieli D (2005) Quantitative trait Loci analysis using the false discovery rate. Genetics 171:783–790
5. Bini C, Ceccardi S, Ferri G et al (2005) Development of a heptaplex PCR system to analyse X-chromosome STR loci from five Italian population samples. A collaborative study. Forensic Sci Int 153:231–236
6. Budowle B, Shea B, Niezgoda S, Chakraborty R (2001) CODIS STR loci from 41 sample populations. J Forensic Sci 46:453–489
7. Carvalho-Silva DR, Pena SD (2000) Molecular characterization and population study of an X chromosome homolog of the Y-linked microsatellite DYS391. Gene 247:233–240
8. Desmarais D, Zhong Y, Chakraborty R, Perreault C, Busque L (1998) Development of a highly polymorphic STR marker for identity testing purposes at the human androgen receptor gene (HUMARA). J Forensic Sci 43:1046–1049
9. Edelman J, Hering S, Michael M et al (2001) 16 X-chromosome STR loci frequency data from a German population. Forensic Sci Int 124:215–218
10. Edelman J, Szibor R (2003) The X-linked STRs DXS7130 and DXS6803. Forensic Sci Int 136:73–75
11. Freeman S (1979) The Pasiegos. Chicago University Press, Chicago
12. Henke L, Aaspollu A, Biondo R, Budowle B, Drobnic K, van Eede PHeal (2003) Evaluation of the STR typing kit Powerplex 16 with respect to technical performance and population genetics: a multicenter study. In: Brinkmann B, Carracedo A (eds) Evaluation of the STR typing kit Powerplex 16 with respect to technical performance and population genetics: a multicenter study. Elsevier, Amsterdam, pp 789–794
13. Kaessmann H, Zollner S, Gustafsson AC et al (2002) Extensive linkage disequilibrium in small human populations in Eurasia. Am J Hum Genet 70:673–685
14. Laan M, Wiebe V, Khusnutdinova E, Remm M, Paabo S (2005) X-chromosome as a marker for population history: Linkage disequilibrium and haplotype study in Eurasian populations. Eur J Hum Genet 13:452–462
15. Moure A, Suárez M (1995) De la Montaña a Cantabria. La construcción de una comunidad autónoma. Publicaciones de la Universidad de Cantabria, Santander
16. Poetsch M, Petersmann H, Repenning A, Lignitz E (2005) Development of two pentaplex systems with X-chromosomal STR loci and their allele frequencies in a northeast German population. Forensic Sci Int 155:71–76
17. Poetsch M, Sabule A, Petersmann H, Volksone V, Lignitz E (2006) Population data of 10 X-chromosomal loci in Latvia. Forensic Sci Int 157:206–209
18. Robino C, Giolitti A, Gino S, Torre C (2006) Development of two multiplex PCR systems for the analysis of 12 X-chromosomal STR loci in a northwestern Italian population sample. Int J Legal Med 120:315–318
19. Ross MT, Bentley DR, Tyler-Smith C (2006) The sequences of the human sex chromosomes. Curr Opin Genet Develop 16:213–218
20. Schaffner SF (2004) The X chromosome in population genetics. Nat Rev Genet 5:43–51
21. Scozzari R, Cruciani F, Malaspina P et al (1997) Differential structuring of human populations for homologous X and Y microsatellite loci. Am J Hum Genet 61:719–733
22. Service S, Deyoung J, Karayuorgou M, Roos JL, Pretorious H, Bedoya Geal (2006) Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. Nat Genet 38:556–560
23. Sobrino B, Brion M, Carracedo A (2005) SNPs in forensic genetics: a review on SNP typing methodologies. Forensic Sci Int 154:181–194
24. Szibor R, Edelman J, Hering S et al (2003) Cell line DNA typing in forensic genetics—the necessity of reliable standards. Forensic Sci Int 138:37–43
25. Szibor R, Hering S, Kuhlisch E et al (2005) Haplotyping of STR cluster DXS6801–DXS6809–DXS6789 on Xq21 provides a powerful tool for kinship testing. Int J Legal Med 119: 363–369
26. Szibor R, Krawczak M, Hering S, Edelman J, Kuhlisch E, Krause D (2003) Use of X-linked markers for forensic purposes. Int J Legal Med 117:67–74
27. Szibor R, Lautsch S, Plate I, Beck N (2000) Population data on the X chromosome short tandem repeat locus HumHPRTB in two regions of Germany. J Forensic Sci 45:231–233
28. Weir BS (1996) Genetic data analysis II. Sinauer, Sunderland
29. Zarrabeitia MT, Alonso A, Martin J et al (2006) Study of six X-linked tetranucleotide microsatellites: population data from five Spanish regions. Int J Legal Med 120:147–150
30. Zarrabeitia MT, Amigo T, Sañudo C, Martinez MA, Riancho JA (2002) Sequence structure and population data of two novel X-linked markers: DXS7423 and DXS8377. Int J Legal Med 116:368–371
31. Zarrabeitia MT, Amigo T, Sañudo C, Zarrabeitia A, González-Lamuño D, Riancho JA (2002) A new pentaplex system to study short tandem repeat markers of forensic interest on X chromosome. Forensic Sci Int 129:85–89
32. Zarrabeitia MT, Riancho JA, Lareu MV, Leyva-Cobian F, Carracedo A (2003) Significance of micro-geographical population structure in forensic cases: a bayesian exploration. Int J Legal Med 117:302–305