Length and repeat-sequence variation in 58 STRs and 94 SNPs in two Spanish populations.

Ferran Casals[a], Roger Anglada[a], Núria Bonet[a], Raquel Rasal[a], Kristiaan J. van der Gaag[b], Jerry Hoogenboom[b], Neus Solé-Morata[c], David Comas[c], Francesc Calafell[c]*

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

[b] Division of Biological Traces, Netherlands Forensic Institute, Laan van Ypenburg 6, 2497 GB, The Hague, The Netherlands

[c] Institut de Biologia Evolutiva (UPF-CSIC), Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, 08003 Barcelona, Catalonia, Spain.

Corresponding author: Dr. Francesc Calafell, Institut de Biologia Evolutiva (UPF-CSIC), Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, 08003 Barcelona, Catalonia, Spain. Tel.: +34 933160842; fax: +34 933160901. E-mail address: francesc.calafell@upf.edu.

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 88 Spanish Roma (Gypsy) samples and 143 Catalans. Since this platform is based in massive parallel sequencing, we have used simple R scripts to uncover the sequence variation in the repeat region. Thus, we have found, across 58 STRs, 541 length-based alleles, which, after considering repeat-sequence variation, became 804 different alleles. All loci in both populations were in Hardy-Weinberg equilibrium. F_{ST} between both populations was 0.0178 for autosomal SNPs, 0.0146 for autosomal STRs, 0.0101 for X-STRs and 0.1866 for Y-STRs. Combined a priori statistics showed quite large; for instance, pooling all the autosomal loci, the a priori probabilities of discriminating a suspect become 1-(2.3×10^{-70}) and 1-(5.9×10^{-73}), for Roma and Catalans respectively, and the chances of excluding a false father in a trio are 1-(2.6×10^{-20}) and 1-(2.0×10^{-21}).

Keywords: Massive Parallel Sequencing; Repeat Sequence-Based Alleles; Roma; Catalans

Introduction

The use of massive parallel sequencing (MPS) is gaining traction in forensic genetics. Although it represents adopting a significantly more complex and expensive technology than capillary electrophoresis (CE), its numerous advantages cannot be ignored. First and foremost, the number of markers, be them STRs or SNPs, that can be simultaneously analyzed is much greater (see details for some commercial kits below), both types of markers can be analyzed simultaneously, and, since in MPS (and unlike in CE), amplicon lengths can freely overlap, many amplicons can be redesigned to a desirable shorter length. Moreover, while preserving the legacy STRs that have been used in the last decades in casework and databanking, MPS allows extracting sequence diversity in alleles that are otherwise isometric and thus seen as equal by CE, possibly incrementing their informativeness.

Several MPS-based amplification kits are already in the market, such as the Applied Biosystems™ Precision ID NGS System, which includes panels for Ancestry (165 SNPs), Identity (124 SNPs), mtDNA whole genome, and STRs (32 STRs plus the amelogenin indel) (https://www.thermofisher.com/content/dam/LifeTech/Documents/PDFs/HID-Precision-ID-System-Brochure.pdf) and Illumina ForenSeq™ [1–7], with Primer Mix A comprising 27 autosomal STRs, 24 Y-STRs, 7 X-STRs and 94 SNPs, and Primer Mix B containing the former plus 22 phenotype-informative SNPs and 56 ancestry-informative SNPs.

Here we report allele and haplotypes frequencies for the Illumina ForenSeq™ Primer Mix A loci, with particular emphasis in sequence allele variation, in two populations from Spain: NE Spanish and Spanish Roma. To the best of our knowledge, this is the first characterization of any European population for this extensive set of 58 STRs and 94 SNPs.

Materials and Methods

Samples

DNA was obtained from saliva in 141 males and two females born in Catalonia of Spanish ancestry (see further details in [8]). Eighty-eight self-identified unrelated Spanish Roma (53 women and 35 men), were sampled in Barcelona, Sant Adrià del Besòs (Barcelona), and Palma. DNA was extracted using a standard organic method with proteinase K digestion, followed by phenol–chloroform extraction. DNA was quantified using Picogreen. This project was reviewed and approved by the Institutional Review Board of the Comitè Ètic d'Investigació Clínica-Institut Municipal d'Assistència Sanitària (CEIC-IMAS) in Barcelona.

Sequencing and allele calling

The Illumina ForenSeq™ Primer Mix A loci were sequenced according to the manufacturer's protocol. Sample volume for amplification and subsequent library preparation was 5µl, at a DNA concentration of 0.2 ng/µl. The pooled libraries were sequenced in a 351 x 31 cycles run with the MiSeq FGx™ instrument following the supplier's protocol. We performed three sequencing runs, with 67, 94, and 94 samples, plus the manufacturer-supplied positive and negative controls in each run. It is worth mentioning that the first run was actually the first ever run in this instrument. Quality metrics (which were always within the boundaries defined by the manufacturer) are shown in Table 1. The negative controls did not yield any result; the positive control failed to provide a genotype for SNP rs1736442 in run 1 and for rs1357617 in run 3.

	Run 1	Run 2	Run 3
Number of samples	67	94	94
Cluster density (k/mm^2)	848	1619	1102
Clusters passing filter	95.22%	87.15%	92.88%
Phasing	0.178%	0.167%	0.178%
Pre-phasing	0.045%	0.076%	0.078%

Table 1. Quality metrics of the three MiSeq FGx™ Illumina ForenSeq™ runs used in this study

Sequencing results were analyzed with the Universal Analysis Software (UAS) provided by the manufacturer. The analytical threshold was set at 0, which means that the hard-coded default of 11 reads to consider a sequence as a potential allele was in effect. The interpretation threshold, that is, the fraction over the total number of reads of the most frequent sequence in an STR or SNP that would trigger the presentation as a sequence that cannot be described by stutter or sequence-based noise, was set to 2.5%. Overall, these settings had the effect to increase the number of sequences that were presented to the user for visual inspection, which was particularly useful in some cases in which low coverage and heterozygote imbalance were both present. Default values were used for the locus-specific stutter thresholds.

Data analysis

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/) and compared against a set of reference sequences we built iteratively from the sequencing results. Thus, we could uncover much more sequence diversity than that provided by the Forenseq UAS interface, which only highlights sequence variants when they are found in isometric heterozygotes; sequence variants in non-duplicated Y-STRs or in X-STRs in males are never highlighted. *IFator* and *YIFator* also produce repeat sequence-based (RSB) allele frequencies; *IFator* provides a priori statistics such as expected heterozygosity, power of discrimination and chance of exclusion; and *YIfator* yields RSB as well as length-based (LB) haplotypes. Note that UAS presents only the sequences of the repeat regions in STRs and not the flanking regions. We devised a shorthand notation for RSB alleles, which is not intended to replace the full nomenclature system proposed by the DNA commission of the ISFG [9]. It consists of the repeat number as provided by UAS (which is as it would have appeared if genotyped by CE), followed by a lowercase letter. If no RSB variation exists, then this letter is *a*. Otherwise, the letters used are meant to capture the structure of the RSB variation. For instance, STR D3S1358 has the general structure TCTA $[TCTG]_x$ $[TCTA]_y$ [9]; 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, or TCTA $[TCTG]_3$ $[TCTA]_{13}$ becomes 17c. Simpler, sporadic non-conformities to a single repeat pattern, or more complex structures were given ad-hoc nomenclatures. As a general criterion, the repeat(s) with the least variants are used for supplemental letters, so as to minimize the number of letters used. The full list of RSB variants and their notation can be found in Supplementary Files 1-3, for autosomal, Y-, and X-STRs.

Hardy-Weinberg and population differentiation tests, as well as F_{ST}, were computed with Arlequin 3.5 [10].

Results

Sequencing results

The Illumina ForenSeq™ Primer Mix A loci were sequenced with the MiSeq FGx™ instrument in 231 samples. Average coverage and heterozygote imbalance (defined as the coverage of the allele with most reads over the total number of reads) are shown in Table 2; detailed results for each locus can be found in Supplementary Files 4 and 5. Average coverage was adequate and similar across locus categories, but across loci it varied by two orders of magnitude, from 29.11 (rs1736442) to 2660.31 (DYS392). Average heterozygote imbalance was always <0.6, with the exception of D5S818 (0.6083) and rs6955448 (0.7045).

	average coverage	coverage range	average het. imbalance	het. imbalance range
Amelogenin	148.98	---	0.5940	---
aSTRs	765.33	155.47 - 1792.60	0.5577	0.5349 - 0.6083
XSTRs	1133.31	41.22 - 1827.61	0.5649	0.5537 - 0.5699
YSTRs	852.32	206.98 - 2660.31	---	----
iSNPs	458.98	29.11 - 1084.00	0.5511	0.5326 - 0.7045

Table 2. Average coverage and heterozygote imbalance for each locus category.

Autosomal STRs

Allele frequencies, heterozygosity, Hardy-Weinberg test results and a priori informativeness statistics for 27 autosomal STRs in Roma and Catalans are presented in Supplementary Files 6 and 7, respectively for LB and RSB alleles. LB genotypes were in Hardy-Weinberg equilibrium ($p>0.05$) in both populations at all STRs after Bonferroni correction. Power of discrimination was $1-(2.9\times10^{-30})$ in Roma and $1-(1.1\times10^{-31})$ in Catalans, which is a reflection of the slightly increased heterozygosity of the general population. Average F_{ST} among both populations was 0.0151, and, after Bonferroni correction, it was significantly different from zero in 11 out of 27 STRs.

RSB variation was detected in 18 out of 27 autosomal STRs, comprising 248 alleles that can be distinguished by sequence but not by length. Of those, less than one third (81) were highlighted by UAS, since it only flags RSB alleles if they are in isometric heterozygotes. Instead, our approach compared each individual's alleles to a reference table. Thus, only in one locus (D4S2408) could UAS uncover all the existent RSB variation, while in five cases, UAS did not detect any of the existing RSB variation. Taking into account RSB variation (Supplementary File 7) genotypes were in Hardy-Weinberg equilibrium (HWE)($p>0.05$). Power of discrimination increased to $1-(1.9\times10^{-33})$ in Roma and $1-(1.9\times10^{-35})$ in Catalans: that is, the average random match probability was 1570 times lower in Roma and 5763 times lower in Catalans with RSB alleles compared to LB alleles. Smaller increases were observed in the chance of excluding a false father in a trio paternity case.

Rare alleles (defined arbitrarily here as those with a frequency <1%) can have an important contribution to solving cases involving distant relatives [11]. In the Roma sample we found 26 LB rare alleles with 24 (27.2%) individuals carrying one, and one (1.1%) individual carrying two; in the larger Catalan sample, we found 48 LB rare alleles, with 42 individuals (29.4%) carrying

one, 7 (4.9%) carrying two, and two (1.4%) carrying three. These figures clearly increased for RSB: Roma individuals carried in total 53 RSB rare alleles, with 39 (44.3%) individuals carrying one rare allele, and 7 (8.0%) individuals carrying two each; in Catalans, 118 rare RSB alleles were found, with only 40 individuals (28.0%) carrying none, and with some individuals carrying as many as six.

In 231 individuals we found 36 RSB alleles not described in the 777 individuals from diverse ethnic backgrounds sequenced in ref. [2], with a maximum frequency of 2.3%. Interestingly, 10 new alleles were found in the Roma and 28 in Catalans, with only two shared by both populations.

X-STRs

Illumina ForenSeq™ contains primers for 7 X-STRs. Allele frequencies, heterozygosity, Hardy-Weinberg test results and F_{ST} values for Catalans and Roma are presented in Supplementary Files 8 and 9 for X-STRs, respectively for LB and RSB alleles. Hardy-Weinberg equilibrium can only be verified in women; since only two women were present in the Catalan sample, we tested for HWE only in Roma, where all seven loci were in HWE both for LB and RSB genotypes. RSB variation was present in five loci, and, across all seven loci, 67 LB alleles but 112 RSB alleles were present. 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, not double heterozygote) females (Supplementary Files 10 and 11). Again, the possibility of identifying RSB alleles implied an increase from a total of 132 different LB haplotypes to 174 RSB haplotypes.

Y-STRs

Out of 24 Y-STRs present in the Illumina ForenSeq™ platform, we could detect RSB alleles in 10 of them (Supplementary Files 12 and 13); overall, the number of alleles increased from 209 LB alleles to 330 RSB alleles. However, RSB variation did not imply an increase in the number of haplotypes (Supplementary File 14); in the case of the Catalan sample, because every male in the sample already carried a different LB haplotype. However, the 33 Roma males for which we obtained complete haplotypes carried 30 different haplotypes (haplotype diversity, 0.9924±0.0104), and men who shared a LB haplotype had also the same RSB haplotype. No haplotypes were shared between Roma and Catalans, and the average F_{ST} was 0.1756 (LB) and 0.1866 (RSB), an order of magnitude larger than for autosomal or X-STRs. 14 Y-chromosome

197 STRs in Illumina ForenSeq™ are shared with AmpFlSTR® *Yfiler*®; for 141 males, AmpFlSTR®
198 *Yfiler*® genotypes were also available [8]. When comparing with the original results, we found
199 13 mismatches out of 1,953 genotypes (21 missing genotypes); however, upon closer
200 inspection, all of them turned out to be clerical or interpretation errors. In particular, two
201 cases concerning DYS437 arose from the fact that the alleles reported by ForenSeq™ were
202 indeed present in the *Yfiler*® electropherogram, but with exceedingly tall peaks (>5000 rfu),
203 which were regarded as noise, and their stutter peaks were mistakenly interpreted as correct.
204
205 *iSNPs*
206 Allele frequencies, a priori statistics, HWE and F_{ST} values for the 94 autosomal identification
207 SNPs in Illumina ForenSeq™ are given in Supplementary File 15. Average expected
208 heterozygosity was close to the maximum possible value of 0.5 both in Roma (0.4544) and
209 Catalans (0.4642); all SNPs were in HWE after Bonferroni correction for multiple testing. The a
210 priori probability of discriminating a suspect was $1-(1.2\times10^{-37})$ in Roma and $1-(3.1\times10^{-38})$ in
211 Catalans; chance of excluding a non-father in a paternity trio was $1-(1.5\times10^{-8})$ and $1-(1.1\times10^{-8})$
212 respectively. When combining all the autosomal loci in Illumina ForenSeq™, these a priori
213 statistics take astounding values: the a priori probabilities of discriminating a suspect become
214 $1-(2.3\times10^{-70})$ and $1-(5.9\times10^{-73})$, and the chances of excluding a false father are $1-(2.6\times10^{-20})$ and
215 $1-(2.0\times10^{-21})$.
216
217 *Comparison with reference populations*
218 Novroski et al. [2] reported the STR allele frequencies in Forenseq™ primer mix A for USA
219 reference populations (African Americans, Asian Americans, European Americans and
220 Hispanics). We have extracted the RSB allele variation in ref. [2] and compared it with the
221 allele frequencies in our own samples (Supplementary File 16). We have also computed F_{ST}
222 (Table 4 and Supplementary File 17). As seen in Table 3, Catalans have very low F_{ST} values with
223 European Americans, and are in fact, for this set of loci, closer to them than to the Spanish
224 Roma.
225

	CAT-ROM	CAT-AFA	CAT-ASN	CAT-CAU	CAT-HIS	ROM-AFA	ROM-ASN	ROM-CAU	ROM-HIS
aSTRs	0.0146	0.0276	0.0273	0.0043	0.0188	0.0331	0.0289	0.0155	0.0210
XSTRs	0.0102	0.0234	0.0322	0.0027	0.0202	0.0271	0.0279	0.0090	0.0173
YSTRs	0.1852	0.1127	0.1484	0.0118	0.0292	0.1677	0.1525	0.1525	0.1718

226

Table 3. Average F_{ST} values by type of locus between Catalans (CAT), Spanish Roma (ROM) and USA reference populations: African Americans (AFA), Asian Americans (ASN), European Americans (CAU), and Hispanics (HIS) [2]

Discussion

We have genotyped two population samples, Roma and Catalans, with the Illumina ForenSeq™ (Primer Mix A); to the best of our knowledge, this is one of the first studies to report allele frequencies and a priori statistics with this platform, outside the reference USA populations described in Novroski et al. [2], who focused in the STRs and did not include the SNPs in this platform. By using MPS, Illumina ForenSeq™ allows accessing not only the allele length of STRs, but their actual sequence. Yet, this information is not easily retrievable from the UAS user interface provided, which only flags repeat sequence variants if found in isometric heterozygotes. Novroski et al. [2] applied a sophisticated bioinformatic approach, which consisted in bypassing UAS, retrieving all the sequences generated in the Illumina ForenSeq™ run, and feeding them to the Strait Razor bioinformatic program [12,13], which produces both RSB and flanking region variants. Considering that not all the forensic laboratories would have the bioinformatic expertise for this approach, we devised a set of simple R scripts than can be run in a variety of operating systems, and that call RSB alleles and compute allele frequencies and a priori statistics. While not covering the whole of sequence variation, they have shown to provide a significant increase in information compared to LB variants.

A consequence of working with RSB variation is that a nomenclature is called for. At the time of writing this work, no such official nomenclature has been adopted. Therefore, we filled a practical need with a relative simple and partly systematic heuristic, which tried to capture both the legacy LB nomenclature and the underlying structure of RSB variation (compare with the semi-random notation used in [2]). In many cases, new, yet undiscovered RSB variants may conform to this pattern (see the example in the Methods section), and a notation to a new allele can be easily given. In other cases, new patterns of variation may emerge, which would then be arbitrarily named. Since this would cause that different laboratories would adopt different names or naming conventions, we must stress that our system should be considered just as shorthand to report our results and not meant to be generalized.

RSB variation, when compared to LB, increases marginally the average STR heterozygosity (Figure 1), yet it increases by several orders of magnitude the overall informativity of this set of STRs. As a general trend, genetic variance between individuals increases, and that among

262 population decreases, as observed in the slight decrease in F_{ST} values between Roma and
263 Catalans. We only observed one case in which a frequent LB allele split into different RSB
264 alleles in each of these two populations: DYS448*19 was found at frequencies 55.9% in Roma
265 and 50.7% in Catalans; it split into two RSB variants, 19a and 19b; the former was present in
266 38.2% of Roma Y chromosomes but was absent in the Catalans, while 19b had frequencies
267 17.7% in Roma and 50.7% in the Catalans. This was then the only case in which RSB allele
268 frequencies implied a substantial increase in F_{ST}, from 0.0021 to 0.1580. Obviously,
269 comparisons between more distantly related populations may reveal more population-specific
270 RSB alleles.

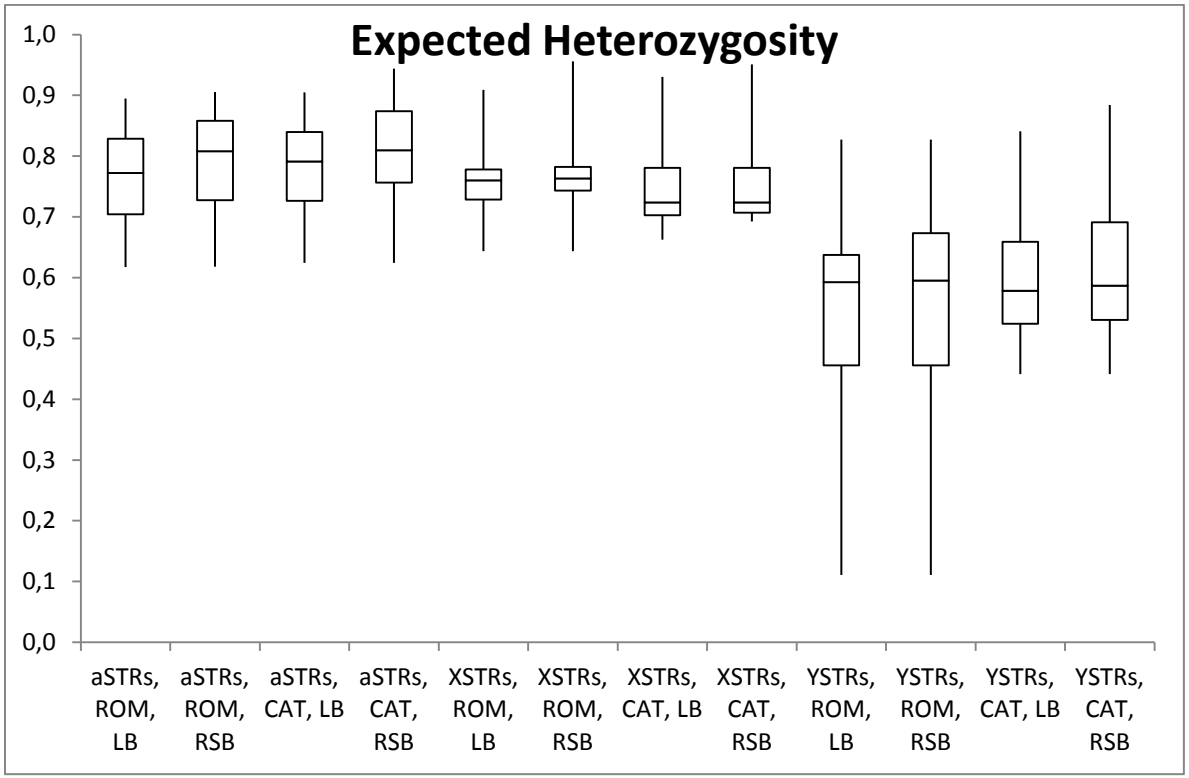

Figure 1. Box plot of the expected heterozygosity by type of variation (length- vs. repeat-sequence based), type of locus (autosomal STRs, X-STRs, Y-STRs), and population. Boxes represent the first and third quartiles; the horizontal line is the median. The whiskers reach to the minimum and maximum values.

Historically, Catalonia has received and integrated migrants from Southern France, Northern Italy, and particularly elsewhere in Spain, making it an open Western European population in genetic terms. Spanish Roma are the product of their Indian origin, travel through the Middle East and Balkans, and admixture with the Spanish host population, while maintaining a certain degree of inbreeding [14,15]. These different population histories are reflected in the diversity patterns found in the Illumina ForenSeq™ loci. Expected heterozygosity was slightly higher in Catalans than in Roma for autosomal STRs (0.8072 vs. 0.7901) and SNPs (0.4622 vs. 0.4544), and for Y-STRs (0.6222 vs. 0.5418), while it was slightly lower in X-STRs (0.7632 vs. 0.7735). Average F_{ST} was 0.0178 for autosomal SNPs, 0.0146 for autosomal STRs, 0.0101 for X-STRs and 0.1866 for Y-STRs. This pattern of increased heterozygosity in Roma in X-STRs but decreased F_{ST} is compatible with previous reports of gene flow into Roma being biased towards women [16,17].

As stated above, RSB variation expands the number of different alleles present even in small population samples such as those that we report. Then, a sufficiently complete description of the RSB diversity in this set of loci may necessitate larger datasets from a comprehensive sample of human populations. This will result from the combined efforts of different laboratories, which undoubtedly will be harmonized when a common nomenclature is adopted.

Acknowledgements

We want to thank the hundreds of volunteers who made this work possible. Marc Tormo (Genomics and Scientific IT Core Facilities, UPF) provided technical support for the computational analyses. Funding was provided by the Spanish Agencia Estatal de Investigación (AEI) and Fondo Europeo de Desarollo Regional (FEDER) (grant CGL2016-75389-P), and by Agència de Gestió d'Ajuts Universitaris i de la Recerca (Generalitat de Catalunya) grant 2014 SGR 866.

References

[1] R.S. Just, L.I. Moreno, J.B. Smerick, J.A. Irwin, Performance and concordance of the ForenSeq™ system for autosomal and Y chromosome short tandem repeat sequencing of reference-type specimens, Forensic Sci. Int. Genet. 28 (2017) 1–9. doi:10.1016/j.fsigen.2017.01.001.

[2] 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. doi:10.1016/j.fsigen.2016.09.007.
[3] A.C. Jäger, M.L. Alvarez, C.P. Davis, E. Guzmán, Y. Han, L. Way, P. Walichiewicz, D. Silva, N. Pham, G. Caves, J. Bruand, F. Schlesinger, S.J.K. Pond, J. Varlaro, K.M. Stephens, C.L. Holt, Developmental validation of the MiSeq FGx Forensic Genomics System for Targeted Next Generation Sequencing in Forensic DNA Casework and Database Laboratories, Forensic Sci. Int. Genet. 28 (2017) 52–70. doi:10.1016/j.fsigen.2017.01.011.
[4] J.D. Churchill, S.E. Schmedes, J.L. King, B. Budowle, Evaluation of the Illumina(®) Beta Version ForenSeq™ DNA Signature Prep Kit for use in genetic profiling., Forensic Sci. Int. Genet. 20 (2016) 20–9. doi:10.1016/j.fsigen.2015.09.009.
[5] C. Xavier, W. Parson, Evaluation of the Illumina ForenSeq™ DNA Signature Prep Kit – MPS forensic application for the MiSeq FGx™ benchtop sequencer, Forensic Sci. Int. Genet. 28 (2017) 188–194. doi:10.1016/j.fsigen.2017.02.018.
[6] 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. doi:10.1016/j.fsigen.2016.05.008.
[7] 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™ DNA Signature Prep Kit STR and SNP loci in Yavapai Native Americans, Forensic Sci. Int. Genet. 28 (2017) 146–154. doi:10.1016/j.fsigen.2017.02.014.
[8] 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. doi:10.1038/ejhg.2015.14.
[9] 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. de Knijff, N. Morling, M. Prinz, P.M. Schneider, C. Van 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. doi:10.1016/j.fsigen.2016.01.009.
[10] 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.
[11] 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). doi:10.1002/elps.201600180.
[12] D.H. Warshauer, D. Lin, K. Hari, R. Jain, C. Davis, B. LaRue, J.L. King, B. Budowle, STRait Razor: A length-based forensic STR allele-calling tool for use with second generation sequencing data, Forensic Sci. Int. Genet. 7 (2013) 409–417. doi:10.1016/j.fsigen.2013.04.005.
[13] D.H. Warshauer, J.L. King, B. Budowle, STRait Razor v2.0: The improved STR Allele Identification Tool – Razor, 2015. doi:10.1016/j.fsigen.2014.10.011.
[14] I. Mendizabal, O. Lao, U.M. Marigorta, A. Wollstein, L. Gusmão, V. Ferak, M. Ioana, A. Jordanova, R. Kaneva, A. Kouvatsi, V. Kučinskas, H. Makukh, A. Metspalu, M.G. Netea, R. de Pablo, H. Pamjav, D. Radojkovic, S.J.H. Rolleston, J. Sertic, M. Macek, D. Comas, M. Kayser, Reconstructing the population history of European Romani from genome-wide data., Curr. Biol. 22 (2012) 2342–9. doi:10.1016/j.cub.2012.10.039.
[15] I. Mendizabal, C. Valente, A. Gusmao, C. Alves, V. Gomes, A. Goios, W. Parson, F. Calafell, L. Alvarez, A. Amorim, L. Gusmao, D. Comas, M.J. Prata, Reconstructing the

Indian origin and dispersal of the European Roma: a maternal genetic perspective, PLoS One. 6 (2011) e15988.

[16] D. Gresham, B. Morar, P.A. Underhill, G. Passarino, A.A. Lin, C. Wiser, D. Angelicheva, F. Calafell, P.J. Oefner, P. Shen, I. Tournev, R. De Pablo, V. Kuĉinskas, A. Perez-Lezaun, E. Marushiakova, V. Popov, L. Kalaydjieva, Origins and divergence of the Roma (Gypsies), Am. J. Hum. Genet. 69 (2001). doi:10.1086/324681.

[17] A. Gusmão, L. Gusmão, V. Gomes, C. Alves, F. Calafell, A. Amorim, M.J. Prata, A perspective on the history of the iberian gypsies provided by phylogeographic analysis of Y-chromosome lineages, Ann. Hum. Genet. 72 (2008). doi:10.1111/j.1469-1809.2007.00421.x.

Supplementary File 1. Nomenclature of RSB alleles in autosomal STRs. The correspondence with the nomenclature in [2] is also given.

Supplementary File 2. Nomenclature of RSB alleles in Y-STRs. The correspondence with the nomenclature in [2] is also given.

Supplementary File 3. Nomenclature of RSB alleles in X-STRs. The correspondence with the nomenclature in [2] is also given.

Supplementary File 4. Average sequencing coverage by locus.

Supplementary File 5. Average heterozygote imbalance (expressed as coverage of the allele with most reads over total number of reads) by locus.

Supplementary File 6. Length-based allele frequencies of autosomal STRs in Roma and Catalans. Sample size, number of different alleles, observed heterozygosity, expected heterozygosity, p-value of the Hardy-Weinberg test, a priori power of discrimination and chance of exclusion, and F_{ST} between Roma and Catalans and its p-value are given for each locus.

Supplementary File 7. Repeat sequence-based allele frequencies of autosomal STRs in Roma and Catalans. Sample size, number of different alleles, observed heterozygosity, expected heterozygosity, p-value of the Hardy-Weinberg test, a priori power of discrimination and chance of exclusion, and F_{ST} between Roma and Catalans and its p-value are given for each locus.

Supplementary File 8. Length-based allele frequencies of X-STRs in Roma and Catalans. Sample size, number of different alleles, expected heterozygosity, p-value of the Hardy-Weinberg test, and F_{ST} between Roma and Catalans and its p-value are given for each locus.

Supplementary File 9. Repeat sequence-based allele frequencies of X-STRs in Roma and Catalans. Sample size, number of different alleles, expected heterozygosity, p-value of the Hardy-Weinberg test, and F_{ST} between Roma and Catalans and its p-value are given for each locus.

Supplementary File 10. Length-based haplotype frequencies of X-STRs in Roma and Catalans. Sample size, number of different halotypes, expected heterozygosity, p-value of the Hardy-Weinberg test, and F_{ST} between Roma and Catalans and its p-value are given for each locus.

Supplementary File 11. Repeat sequence-based haplotype frequencies of X-STRs in Roma and Catalans. Sample size, number of different halotypes, expected heterozygosity, p-value of the Hardy-Weinberg test, and F_{ST} between Roma and Catalans and its p-value are given for each locus.

Supplementary File 12. Length-based allele frequencies of Y-STRs in Roma and Catalans. Sample size, number of different alleles, expected heterozygosity, and F_{ST} between Roma and Catalans and its p-value are given for each locus.

Supplementary File 13. Repeat sequence-based allele frequencies of Y-STRs in Roma and Catalans. Sample size, number of different alleles, expected heterozygosity, and F_{ST} between Roma and Catalans and its p-value are given for each locus.

Supplementary File 14. Repeat sequence-based Y-STR haplotype frequencies in Roma and Catalans.

Supplementary File 15. Allele frequencies in Roma and Catalans for 94 autosomal SNPs. Sample size, observed heterozygosity, expected heterozygosity, p-value of the Hardy-Weinberg test, a priori power of discrimination and chance of exclusion, and F_{ST} between Roma and Catalans and its p-value are given for each locus.

Supplementary File 16. Repeat sequence-based allele frequencies for autosomal, X-, and Y-STRs in Roma (ROM) and Catalans (CAT), and in African Americans (AFA), Asian Americans (ASN), European Americans (CAU) and Hispanics (HIS) [2].

Supplementary File 17. F_{ST} values by locus between Catalans (CAT), Spanish Roma (ROM) and USA reference populations: African Americans (AFA), Asian Americans (ASN), European Americans (CAU), and Hispanics (HIS) [2].