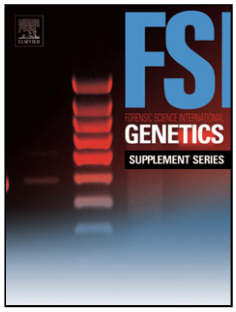
Article

Sequence-Based Saudi Population Data for The SE33 Locus

Alsafiah, Hussain M., Khubrani, Yahya M., Hadi, Ss and Goodwin, William H

Alsafiah, Hussain M., Khubrani, Yahya M., Hadi, Ss ORCID: 0000-0002-2994-3083 and Goodwin, William H ORCID: 0000-0002-3632-3552 (2019) Sequence-Based Saudi Population Data for The SE33 Locus. Forensic Science International: Genetics Supplement Series, 7 (1). pp. 9-11. ISSN 1875-1768

Journal Pre-proof



Sequence-Based Saudi Population Data for The SE33 Locus

Hussain M. Alsafiah, Yahya M. Khubrani, Hadi Sibte, William H. Goodwin

PII: S1875-1768(19)30187-8

DOI: https://doi.org/10.1016/j.fsigss.2019.09.004

Reference: FSIGSS 1509

To appear in: *Forensic Science International: Genetics Supplement Series*

Received Date: 17 September 2019

Revised Date: 17 September 2019

Accepted Date: 17 September 2019

Please cite this article as: Alsafiah HM, Khubrani YM, Sibte H, Goodwin WH, Sequence-Based Saudi Population Data for The SE33 Locus, *Forensic Science International: Genetics Supplement Series* (2019), doi: https://doi.org/10.1016/j.fsigss.2019.09.004

Sequence-Based Saudi Population Data for The SE33 Locus

Alsafiah, Hussain M.[1, 2]; Khubrani, Yahya M.[1,3]; Sibte Hadi [2]; Goodwin, William H.* [2]

[1] Forensic Genetics Laboratory, General Administration of Criminal Evidences, Public Security, Ministry of Interior, Kingdom of Saudi Arabia.
[2] School of Forensic and Applied Sciences, University of Central Lancashire, Preston, United Kingdom.
[3] Department of Genetics & Genome Biology, University of Leicester, Leicester, United Kingdom.

*Corresponding Author: whgoodwin@uclan.ac.uk

Abstract

A set of 87 reference samples collected from the population of Saudi Arabia were sequenced using the ForenSeq™ DNA Signature Prep Kit on a MiSeq FGx™. The FASTQ files contain the sequences of the SE33 STR, but are not reported by the ForenSeq™ Universal Analysis Software (UAS). The STRait Razor software was used to recover and to report SE33 sequence-based data for the Saudi population. Ninety-six sequence-based alleles were recovered, most of which had previously reported motif patterns. Two unreported motif patterns found in three alleles and seven novel allele sequences were reported. We also reported a single discordance between the sequence-based data and the CE data that was due to the presence of a common TTTT deletion. SE33 had 130% more sequence-based alleles; the highest number of observed sequence variants were in alleles 27.2 and 30.2, which each had 7 sequence variants. The statistical parameters emphasize the usefulness of using the sequence-based data.

keywords:

Massively parallel sequencing, Saudi Arabia, SE33.

Introduction

Massively Parallel Sequencing (MPS) systems are now being adopted in many forensic laboratories generating detailed sequence data for different type of markers simultaneously. The ForenSeq™ DNA Signature Prep Kit allows sequencing >150 (Primer Mix A) or >230 markers (Primer Mix B) where users can decide which primer mix will be used.

By utilising the MiSeq FGx™ and ForenSeq™ DNA Signature Prep Kit, the ForenSeq™ Universal Analysis Software (UAS), reports 27 autosomal STRs (aSTRs) along with other commonly used markers (Y-STRs, X-STRs, and SNPs). Although SE33 is included, it is not reported by the UAS.

SE33 is the most polymorphic well-characterised STR [1] which makes it valuable for forensic applications. Previous studies have demonstrated that sequence-based data of SE33 presents significantly more observed alleles compared to CE systems. A recent study has classified the SE33 repeat motifs into 34 types (A0, A1, A2…. to D3) based on the structure of the repeat region, eleven of which had >1% frequency in the tested populations [2].

The aim of this study was to provide Saudi sequence-based data for the SE33 locus. This included a concordance study with the GlobalFiler™ PCR Amplification Kit.

Materials and Methods

Eighty-seven reference samples from the Saudi population, which were already profiled with the GlobalFiler® kit [3], were sequenced in the study. Using ForenSeq™ DNA Signature Prep Kit (Primer Mix A), libraries were prepared for sequencing following the manufacturer's guidelines except that the volume of the pooled normalised library (PNL) was increased from 7 µl to 12 µl. Sequencing was carried out using a MiSeq FGx™ following the manufacturer's guidelines.

The STRait Razor v3.0 (SR) [4] , was used to recover the SE33 sequences from the FASTQ files after modifying the configuration file by adding the 5' and 3' anchors and motif sequence provided in [2]. All sequences with ≥ 10 reads (depth of coverage (DoC)) and heterozygous sequences that showed ≥ 20% of allele coverage ratio (ACR), were recovered automatically by the software. Sequences that showed less than 20% ACR were recovered manually.

For the concordance study, the sequenced-based data was compared to CE data and Sanger sequencing was used to confirm a discordant result as previously described in [5]. The novelty of a motif pattern or of an allele sequence was assessed based on those motifs and sequences reported in [2] and in STRBase [6].

Allele frequencies, matching probability (MP), and power of exclusion (PE) were calculated using the GenAlEx 6.5 software [7]. The typical paternity index (PI) was calculated using the

equation PI = (h+H)/2h (h=homozygosity and H = Heterozygosity) as described in [8]. Finally, the expected heterozygosity (He), observed heterozygosity (Ho) and the exact test for Hardy-Weinberg Equilibrium (HWE) were calculated using Arlequin v 3.5 [9].

Results and discussion

The SE33 sequences of the 87 samples were recovered, 83 of which were within the designated limits (≥ 10 reads and ≥ 20% ACR), and the rest of samples (four samples) were recovered manually due lower ACR (< 20%). The ACR of heterozygous sequences ranged from 6.5% to 99.4% and showed an average of 58.6%, the four manually typed samples had ACR of 6.5% (alleles 6.3, 31.2), 8.14% for alleles (14,35.2), 12.17% for alleles (13.3,31.2), and 12.8% for alleles (17,34). Among the 87 samples, these samples had the largest size difference between the long and short allele that ranged from 99 bp to 68 bp demonstrating the ACR correlation with the size difference of the heterozygous allele pair.

The total coverage of the SE33 locus in all samples was 53,956 reads and the average DoC of recovered sequences was 742 reads that ranged from 32 to 2196 reads for alleles 31.2 and 6.3 respectively.

The number of observed sequence-based alleles was 130% more (69 alleles) comparing to 30 size-based alleles. Most sequence variants (iso-alleles) were observed in x.2 alleles where alleles 27.2 and 30.2 had the highest number of observed sequences (7 sequence variants/allele).

The SE33 motif patterns of the 69 sequences showed that 66 alleles were within the classification of Borsuk *et al.* (2018) [2] and most of these alleles (53 alleles), as expected, had an A0 or A1 motif. Two new motif patterns were observed in three alleles that are shown in Table 1. Following on from the earlier study we suggest two new motif IDs (D4 & D5). In addition, seven sequences, which fall within the motif classification, but have not been reported before were observed (Table 1).

A single discordance was observed, where the sample had 19,31.2 in the sequence-based data while it had 18,31.2 in the size-based data. The allele 19 had CT $[CTTT]_3$ C **$[CTTT]_{19}$** CT $[CTTT]_3$ CT $[CTTT]_2$ (counted part of the repeat region is in bold) suggesting a deletion of four bp within the flanking region. Examination of the FASTQ file of the sample revealed a [TTTT]

deletion at 88277355_88277358 (GRCh38) when compared to the reference sequence of the locus. This was further investigated by Sanger sequencing and the deletion was confirmed. The deletion was assigned rs369314007 and was found to be associated with the A0 motif [2], which is the motif pattern of allele 19.

The data showed that the heterozygosity was increased from 90.8% (79 heterozygous samples) to 91.9% (80 heterozygous samples), and both data were within the expectations of HWE (*P* value > 0.05).

For the population of Saudi Arabia, the sequence data of the SE33 locus showed that the power of discrimination, power of exclusion and the typical paternity index increased from 99.3% to 99.7%, 89.4% to 93%, and from 5.44 to 6.21 respectively. The figures emphasize the value of using SE33 in forensic applications especially with mixture analysis and in paternity testing.

Conclusion

This study provides sequence-based Saudi population data for the SE33 locus for the first time. As expected, most sequences showed A0 and A1 motif patterns while three sequence-based alleles were not within the classification. Two new motif patterns are reported, and their motif IDs were suggested as D4 and D5. In addition, seven sequences that fall within the classification but have not been reported before, were reported. The discordance event was resolved by Sanger sequencing that showed the presence of the rs369314007 deletion.

Acknowledgement

We would like to thank Professor Mark A. Jobling and Dr. Jon Wetton (University of Leicester) for allowing us to carry out the lab work and analysis of the samples in Alec Jeffreys Forensic Genomics Unit. The study was funded by the Royal Embassy of Saudi Arabia Cultural Bureau in London (UKSAcb).

Conflict of interest

None

References

[1] P. Wiegand, B. Budowle, S. Rand, B. Brinkmann, Forensic validation of the STR systems SE 33 and TC 11, International Journal of Legal Medicine 105 (1993) 315-320.

[2] L.A. Borsuk, K.B. Gettings, C.R. Steffen, K.M. Kiesler, P.M. Vallone, Sequence-based US population data for the SE33 locus, Electrophoresis. 39 (2018) 2694-2701.

[3] H.M. Alsafiah, W.H. Goodwin, S. Hadi, M.A. Alshaikhi, P. Wepeba, Population genetic data for 21 autosomal STR loci for the Saudi Arabian population using the GlobalFiler® PCR amplification kit, Forensic Science International: Genetics 31 (2017) e59-e61.

[4] A.E. Woerner, J.L. King, B. Budowle, Fast STR allele identification with STRait Razor 3.0, Forensic Science International: Genetics 30 (2017) 18-23.

[5] H.M. Alsafiah, A. Iyengar, S. Hadi, W.M. Alshlash, W. Goodwin, Sequence data of six unusual alleles at SE33 and D1S1656 STR Loci, Electrophoresis 39 (2018) 2471-2476.

[6] C. Ruitberg, D. Reeder, J. Butler, STRBase: a short tandem repeat DNA database for the human identity testing community, Nucleic Acids Research 29 (2001) 320-322.

[7] R. Peakall, P.E. Smouse, GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research _an update, Bioinformatics. 28 (2012) 2537-2539.

[8] C. Brenner and J. Morris, Paternity index calculations in single locus hypervariable DNA probes: validation and other studies, In Proceedings for the International Symposium on Human Identification (1989). Promega Corporation, Madison, WI.

[9] 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, Molecular Ecology Resources. 10 (2010) 564-567.

Table 1. Motif patterns of the SE33 locus observed in the samples from Saudi Arabia. A total of 66 allele sequences were within motif patterns classified by Borsuk *et al.* (2018) [2], 53 of which, as expected, had the A0 and A1 motif patterns. Two unreported motif patterns were observed in three alleles and were given D4 and D5 motif IDs. Rows in red indicates novel motifs observed in Saudi population.

Alleles	Motif	Obs.	ID	Novelty
9-22	CT [CTTT]3 C [CTTT]n CT [CTTT]3 CT [CTTT]2	13	A0	Novel sequence (Allele 9)
20.2-33.2	CT [CTTT]2 CCTT C [CTTT]n TT [CTTT]n CT [CTTT]3 CT [CTTT]2	40	A1	Reported in [2]
30.2	CT [CTTT]2 CCTT C [CTTT]n CT [CTTT]n CT [CTTT]3 CT [CTTT]2	1	A3	Reported in [2]
34	CT [CTTT]2 CCTT C [CTTT]n TT [CTTT]n TT [CTTT]n CT [CTTT]3 CT [CTTT]2	1	A7	Novel sequence
35.2	CT [CTTT]2 CCTT C [CTTT]n TT [CTTT]n CT [CTTT]3 CT CTTT	1	A8	Novel sequence
6.3 & 7.3	CT [CTTT]3 [CTTT]n CT [CTTT]3 CT [CTTT]2	2	C2	Novel sequence (Allele 7.3)
13.3	CT [CTTT]3 C [CTTT]n C [CTTT]n [CTTT]3 CT [CTTT]2	1	B2	Novel sequence
18	CT [CTTT]2 C [CTTT]n CT [CTTT]3 CT [CTTT]2	1	B1	Reported in [2]
20.2 & 22.2	CT [CTTT]3 C [CTTT]n CT [CTTT]n CT [CTTT]3 CT [CTTT]2	2	B3	Novel sequence
26.2	CT [CTTT]2 [CCTT]3 C [CTTT]n TT [CTTT]n CT [CTTT]3 CT [CTTT]2	1	B9	Reported in [2]
28.2	CT [CTTT]2 [CCTT]2 C [CTTT]n TT [CTTT]n CT [CTTT]3 CT [CTTT]2	1	A4	Reported in [2]
27.2	CT [CTTT]2 CCTT C [CTTT]n CTGT [CTTT]n TT [CTTT]n CT [CTTT]3 CT [CTTT]2	1	C4	Novel sequence
27.2	CT [CTTT]2 CCTT C [CTTT]n TT [CTTT]n CT TTTT [CTTT]2 CT [CTTT]2	1	D4*	Novel motif
29.2 & 30.2	CT [CTTT]2 CCTT C [CTTT]n CCTT [CTTT]n TT [CTTT]n CT [CTTT]3 CT [CTTT]2	2	D5*	Novel motif
30.2	CT [CTTT]2 C [CTTT]n TT [CTTT]n CT [CTTT]3 CT [CTTT]2	1	B5	Reported in [2]

*The D4 and D5 IDs were suggested to continue the work done by Borsuk *et al.* (2018) [2]