# <PE-AT>Sequence Data of Six Unusual Alleles at SE33 and D1S1656 STR Loci

Hussain Mohammed Alsafiah[1,2], Arati Iyengar[1], Sibte Hadi[3], Waleed M. Alshlash[4], William Goodwin[1]

1 University of Central Lancashire - Forensic and Investigative Sciences
Maudland Building, Preston, Lancashire PR1 2HE
United Kingdom of Great Britain and Northern Ireland

2 Forensic Genetics Laboratory, General Administration of Criminal Evidences, Public Security, Ministry of Interior, Kingdom of Saudi Arabia -
Forensic Genetics Laboratory, General Administration of Criminal Evidences , Riyadh, Riyadh
Saudi Arabia

3 UCLan - Forensic Sciences
JB Firth Building , Preston PR! 2HE
United Kingdom of Great Britain and Northern Ireland

4 Forensic Genetics Laboratory, General Administration of Criminal Evidences, Public Security, Ministry of Interior, Kingdom of Saudi Arabia -
Forensic Genetics Laboratory, General Administration of Criminal Evidences, Riyadh, Riyadh 11343
Saudi Arabia

Corresponding author:
William Goodwin
University of Central Lancashire - Forensic and Investigative Sciences
Maudland Building, Preston, Lancashire PR1 2HE
United Kingdom of Great Britain and Northern Ireland
whgoodwin@uclan.ac.uk
http://orcid.org/0000-0002-3632-3552

Received: 05 01, 2018; Revised: 07 10, 2018; Accepted: 07 10, 2018

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/elps.201800191.

## Abstract

When profiling a reference dataset of 500 DNA samples for the population of Saudi Arabia, using the GlobalFiler® PCR amplification kit, six unusual alleles were detected. At the SE33 locus, four novel alleles were found: 2, 14.3, 20.3, and 38; two alleles, at the D1S1656 locus: 7 and 8, had been previously reported, but no published sequence data was available.

The D1S1656 alleles were sequenced using ForenSeq™ DNA Signature Prep with the MiSeq FGx System (Illumina, USA). As the SE33 is not reported by available Massively Parallel Sequencing (MPS) systems, samples that exhibited the unreported alleles were sequenced using BigDye™ Terminator v3.1 Cycle Sequencing Kit. Here we present the sequence and structure of the previously uncharacterized alleles.


**Keywords**: D1S1656 , Massively Parallel Sequencing , Saudi Arabian population, SE33 , STR variant alleles

<FRONTEND>

## Introduction

The SE33 and D1S1656 loci are highly informative for the population of Saudi Arabia and showed a power of discrimination of 0.993 and 0.970 respectively [1]. The SE33 alleles detailed here have not been reported in STRBase; the D1S1656 alleles 7 and 8 have been reported in the STRBase [2], but no sequence data was available (Figures 1-3).

SE33 is the most polymorphic STR locus that is commonly used in forensic genetics [3]. The landscape of this locus is well characterised, which is divided to three regions: repeat region, “local flanks”, and extended flanks [4]. The sequence structure of the repeat region is based on tetra-nucleotides repeats of [(CTTT) n] and the complexity of the structure increases as alleles become larger [5, 6]. Shorter alleles < 260 bp tend to have the basic repeat unit, and alleles between 263 to 315 bp are typically associated with the presence of one hexa-nucleotide repeat (CTTTTT), within the repeat units; these were classified as type 1 and 2 respectively [5] (Table 1A). Larger alleles (≥ 317 bp) were found to have two hexa-nucleotide repeats (CTTTTT) interrupting the repeat region [6], which will be classified here as type 3 (Table 1A).

The D1S1656 has a compound repeat structure of [(TAGA) n (TAGG)] followed by [(TG) $_5$] in the 3´ Local flank (Table 1B). This locus was added to the European Standard Set (ESS) in 2006 [7] and to the Combined DNA Index System (CODIS) in 2015 [8].

Based on the sequence-based nomenclature guidelines of the International Society for Forensic Genetics (ISFG), the local flank regions showed in Table 1 A and B, are not counted in allele calling system [9].

Previous studies have demonstrated that both loci contain a high level of sequence-based variations. With sequencing, the number of characterized alleles (iso-alleles) increased by 250% with SE33 and by 64% with D1S1656, compared to length-based systems [10, 11]**.**

Although SE33 is already included in the primer mixes of the ForenSeq™ DNA Signature Prep (Illumina, USA) [12], it is not reported by the ForenSeq™ Universal Analysis Software (Illumina, USA). This may be due to the high dropout rate that was observed when analysing the ForenSeq™ data using an independent software [4]. The highly repetitive sequence of the extended flanking regions makes the size of amplicons large that reduces the read quality [10]. In addition, thymine and cytosine represent more than 80% of the forward strand of the SE33 amplicons that adds more challenges to be sequenced [4]. In contrast, the D1S1656 is already included and is reported in Precision ID GlobalFiler MPS Panel (ThermoFisher, USA) and in the ForenSeq™ DNA Signature Prep (Illumina, USA), for example [13, 14] .

The observation of alleles outside the designated windows of an allelic ladder, i.e. allele 2 and 7, may lead to misinterpretation of this allele highlighting the importance of understanding sequence structure. Therefore, this study aimed to provide sequence data for the SE33 alleles using BigDye™ Terminator v3.1 Cycle Sequencing Kit (ThermoFisher, USA) and for the D1S1656 alleles using ForenSeq DNA Signature Prep Kit (Illumina, USA).

## Materials and Methods

For the SE33 alleles, samples that exhibited alleles of interest were amplified in 25 µl volume reactions that contained 12.5 µl of ReddyMix PCR Master Mix (ThermoFisher, USA), 1.25 µl of a 10-µM concentration of each primer, and a total of 2 ng DNA. A primer pair (SE33-1 and SE33-2), as published in [15], was used for the amplification reactions: SE33-1 5´-AAT CTG GGC GAC AAG AGT GA-3´ and SE33-2 5´-ACA TCT CCC CTA CCG CTA TA-3´. This is the same set used by Moller *et al*. [5] and Rolf *et al*. [6], who reported the sequence structure of the repeat region. Samples were amplified using a Veriti™ Thermal Cycler (ThermoFisher, USA): [95 °C / 2 min] [(95 °C / 25 s) (60 °C / 30 s) (72 °C / 40 s)] 30 cycles [72 °C / 5 min].

A 20-cm-long 3% agarose gel was employed to separate target alleles. A total of 25 µl of the PCR products and 10 µl of 100 bp DNA Ladder Plus (NBS-biologicals, UK) were loaded; electrophoresis was at 120 v for 6 h. DNA from the targeted bands was recovered using PureLink™ Quick Gel Extraction Kit (ThermoFisher, USA) following the manufacturer's procedure. DNA concentrations were estimated using Qubit™ dsDNA HS Assay Kit and Qubit® Fluorometer 3.0 (ThermoFisher, USA).

DNA was sequenced directly using BigDye™ Terminator v3.1 Cycle Sequencing Kit (ThermoFisher, USA) following an internally validated 10 µl reaction volume. For each DNA strand, the 10 µl

sequencing reaction contained 0.75 µl of BigDye® Terminator v3.1 Ready Reaction Mix, 1.7 µl 5X Sequencing Buffer, 0.32 µl of 10 µM primer (forward or reverse), and 3-6 ng of DNA. A Veriti™ Thermal Cycler was used for sequencing reaction: [95 °C / 1 min] [(96 °C / 10 s) (50 °C / 5 s) (60 °C / 4 min)] 25 cycles.

Post sequencing purification was carried out by adding 2 µl Shrimp Alkaline Phosphatase (SAP) (ThermoFisher, USA) to 5 µl of sequencing products that was followed by an incubation at 37 °C for 60 min then at 65°C for 15 min as recommended by the manufacturer.

Purified products were prepared for separation by adding 5 µl Hi-Di™ Formamide (ThermoFisher, USA). An ABI 3500 DNA Genetic Analyser, POP-6™ polymer and 50 cm capillary array were employed for separation using the run modules StdSeq50_POP6 and the basecalling protocol BDTv3.1_PA_Protocol-POP6. Sequencing raw data was then analysed by sequencing analysis software v5.4 (ThermoFisher, USA).

ForenSeq™ DNA Signature Prep (primer mix B) and MiSeq FGx System (Illumina, USA) were used to genotype 230 forensic markers including the D1S1656. This was carried out in the Applications Laboratory, Illumina, Cambridge, United Kingdom; following the manufacturer's guidelines. ForenSeq Universal Analysis Software was employed for results analysis by applying the default settings.

## Results and Discussion

The SE33 locus was successfully amplified from samples that exhibited the alleles 2, 14.3, 20.3 and 38. The physical separation of the alleles was successfully achieved when using the 20-cm-long agarose gel (

Figure 4). DNA recovery from the targeted bands yielded adequate concentrations (0.25 ng/µl to 0.78 ng/µl) to achieve successful direct sequencing for all alleles.

Based on size-based system, allele 2 could be due to a complete loss of the repeat region or due to sequence deletion within the flanking regions. However, the presence of a stutter artefact, which is associated with the repetitive regions, suggested sequence deletion in the flanking regions (Figure 1). As expected, sequence data revealed that the allele 2 consisted of 16 repeats with a 60 bp deletion in the extended 3´ flank, which was accompanied with an additional CTTT repeat in the uncounted 5´ Local flank (Table 2). Although, SE33 allele 14.3 showed [CTTT] $_{2}$ C [CTTT] $_{15}$ (allele 17.1), a 14 bp deletion in the extended 3´ flank accompanied with an additional CTTT repeat in the uncounted 5´ Local flank led to the observation of allele 14.3, based on size (Table 2). The alleles 2 and 14.3 showed a T variant at location 6:88277143 in the 5´ Local flank, which represents rs536914220 SNP (C: 97% T: 3%) (Table 2). Allele 14.3 also showed a T variant at location 6:88277260 in the 3´ Local flank, which represents rs1045867314 SNP (no population statistics available in the 1000 Genome data) (Table 2). Allele 20.3 showed three T nucleotides within the repeat sequence that could have occurred due to a C deletion in a single repeat or due to an insertion of three T nucleotides. As expected from larger alleles, allele 38 showed a typical type 3 structure that exhibits two hexanucleotide repeats within the repeat region (Table 2).

Samples that showed alleles 7 and 8 at the D1S1656 was successfully sequenced using the ForenSeq DNA Signature Prep Kit and the MiSeq FGx Forensic System. Both samples showed 100 % concordance at 21 autosomal STRs and DYS391 loci overlapped with the GlobalFiler® PCR amplification kit. Allele 7 showed a typical sequence structure of [(TAGA) $_{6}$ (TAGG) $_{1}$] (

Figure 5 A). However, allele 8 showed [(TAGA) $_{8}$] sequence where the (TAGG) repeat was absent (Figure 2 B). This absence was previously reported in [11, 16], which could be interpreted by the presence of an A variant of rs78443572 SNP (TAGG, G: 73%, A: 27%) [17].

## Conclusion

This study has provided sequence data for six previously uncharacterized alleles. The SE33 alleles 2, 14.3, 20.3 had type 1 structure and allele 38 had Type 3 structure. In addition, based on the sequence-based nomenclature guidelines of the ISFG, the alleles 2 and 14.3 at the SE33 should be called as 16 and 17.1 respectively. The assumption that allele 2 at the SE33 was due to deletion in a flanking region was confirmed by the sequence data. The observation of alleles outside the designated windows of an allelic ladder may lead to misinterpretation of this allele that was resolved by analysing the sequence structure.

**Ethical approval**

Samples tested in this study were collected after receiving ethical approvals from the Security Forces Hospitals Programme in Saudi Arabia and from the Ethics Committee of the University of Central Lancashire, UK.

**Acknowledgement**

We would like to thank Richard Kessell (Verogen, UK) and Sarah Naif (Illumina, UK) for carrying out the Massively Parallel Sequencing at the Illumina Applications Laboratory, Cambridge, UK. This study has been funded by Saudi Arabian Cultural Bureau in London.

**Conflicts of Interest**

The authors have declared no conflict of interest.

# References

[1] Alsafiah, H. M., Goodwin, W. H., Hadi, S., Alshaikhi, M. A., Wepeba, P., *Forensic. Sci. Int. Genet*. 2017, *31*, e59-e61.

[2] Ruitberg, C., Reeder, D., Butler, J., *Nucleic Acids Res.* 2001, *29*, 320-322.

[3] Wiegand, P., Budowle, B., Rand, S., Brinkmann, B., *Int. J. Legal Med.* 1993, *105*, 315-320.

[4] Borsuk, L. A., Gettings, K. B., Steffen, C. R., Kiesler, K. M., Vallone, P. M., *Electrophoresis* 2018, *0*, 1-8.

[5] Moller, A., Brinkmann, B., *Int. J. Legal Med.* 1994, *106*, 262-267.

[6] Rolf, B., Schurenkamp, M., Junge, A., Brinkmann, B., *Int. J. Legal Med.* 1997, *110*, 69-72.

[7] Gill, P., Fereday, L., Morling, N., Schneider, P. M., *Forensic Sci. Int.* 2006, *163*, 155-157.

[8] Hares, D. R. *Forensic. Sci. Int. Genet.* 2015, *17*, 33-34.

[9] Parson, W., Ballard, D., Budowle, B., Butler, J. M. *et al*., *Forensic Sci. Int. Genet.* 2016, *22*, 54-63.

[10] Gettings, K. B., Aponte, R. A., Vallone, P. M., Butler, J. M., *Forensic. Sci. Int. Genet.* 2015, *18*, 118-130.

[11] Gettings, K. B., Kiesler, K. M., Faith, S. A., Montano, E. *et al*., *Forensic. Sci. Int. Genet.* 2016, *21*, 15-21.

[12] Novroski, N. M. M., King, J. L., Churchill, J. D., Seah, L. H., Budowle, B., *Forensic Sci. Int. Genet.* 2016, *25*, 214-226.

[13] Guo, F., Yu, J., Zhang, L., Li, J., *Forensic. Sci. Int. Genet.* 2017, *31*, 135-148.

[14] Wang, Z., Zhou, D., Wang, H., Jia, Z. *et al.*, *Forensic Sci. Int. Genet.* 2017, *31*, 126-134.

[15] Gill, P., Kimpton, C., D'Aloja, E., Andersen, J. F. *et al.*, *Forensic Sci. Int.* 1994, *65*, 51-59.

[16] Kline, M. C., Hill, C. R., Decker, A. E., Butler, J. M., *Forensic. Sci. Int. Genet.* 2011, *5*, 329-332.

[17] The 1000 Genomes Project Consortium, *Nature* 526. 7571, Web (Accessed 13 July 2017), 68–74.

**Figure 1. An electropherogram showing allele 2 at the SE33 locus. The Figure shows an unusual SE33 allele situated under the D7S820 allelic window; this allele was previously demonstrated to belong to the SE33 locus and was called as allele 2 based on its size [1]. This short allele could be due to a complete loss in the repeat region or due to a deletion in the flanking regions. However, the observation of the stutter artefact before this allele (black arrow), which is associated with repetitive regions, suggested a deletion in the flanking regions.**

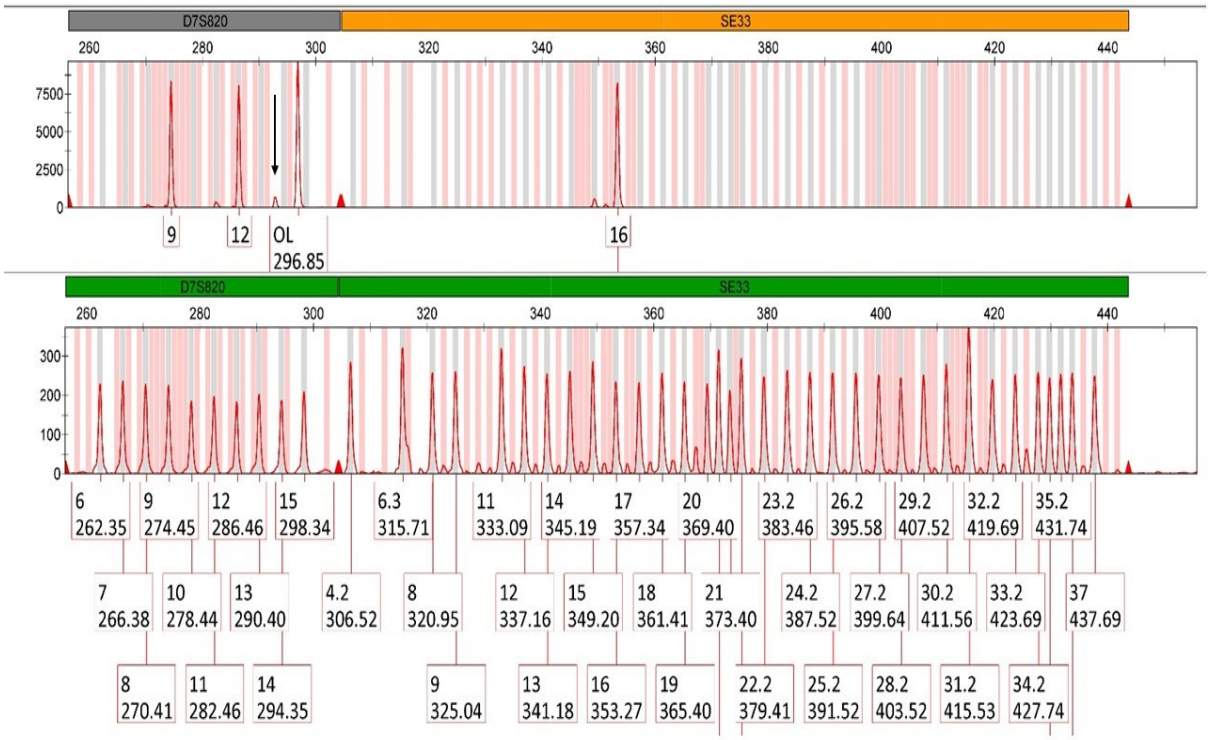


**Figure 2. Electropherograms for three novel alleles at the SE33 locus (A-C) with an allelic ladder for reference (D). These alleles were called based on their sizes as 14.3 (A), 20.3(B), and 38(C).**

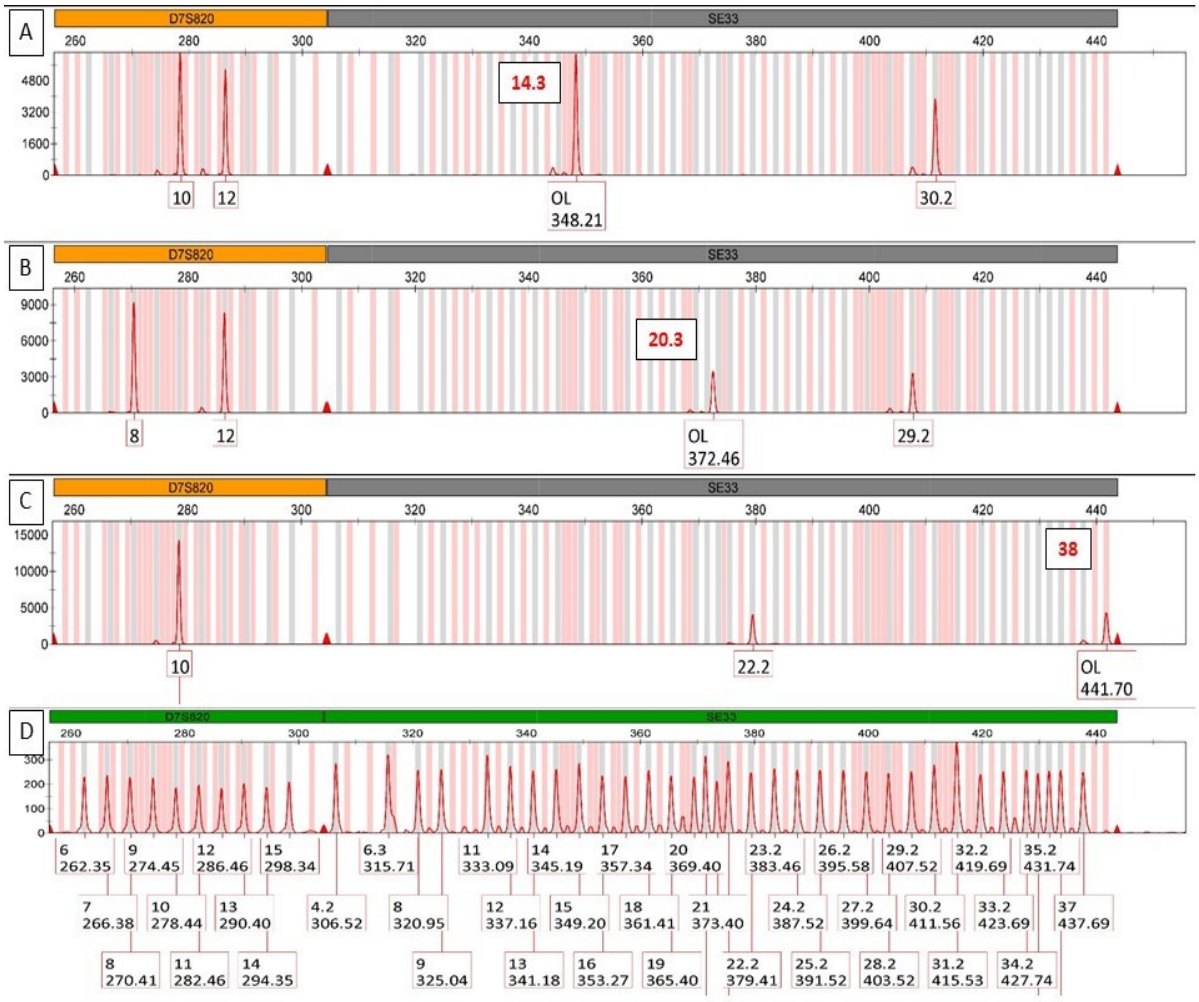



**Figure 3. Electropherograms for the two alleles of the D1S1656 locus (A and B). These allele were called based on their sizes as allele 7 and 8; (C) shows an allelic ladder.**

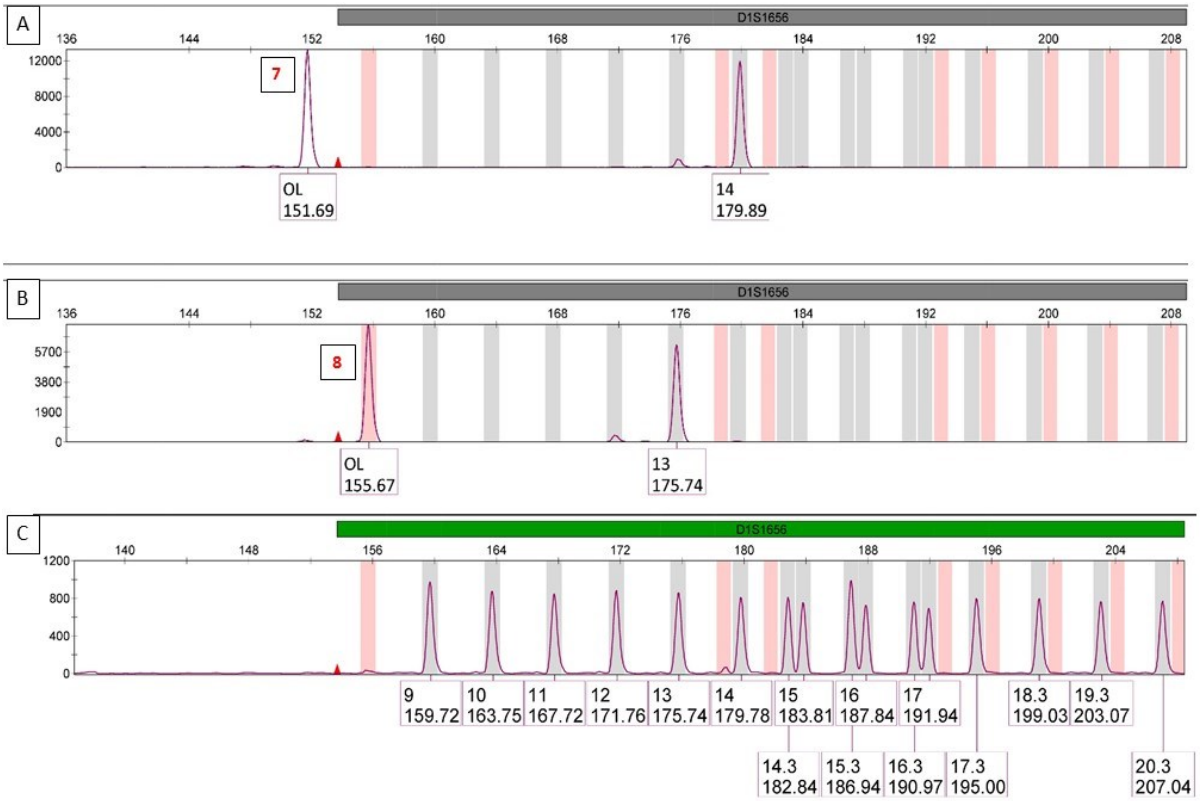


**Figure 4. A 20-cm-long 3% agarose gel with the novel SE33 alleles; from the left side, alleles 20.3, 14.3, 38, 2 and a 100 bp ladder. It shows the separation of alleles 20.3 and 29.2 (35 bp) that could not be achieved with a shorter (10 cm) gel.**

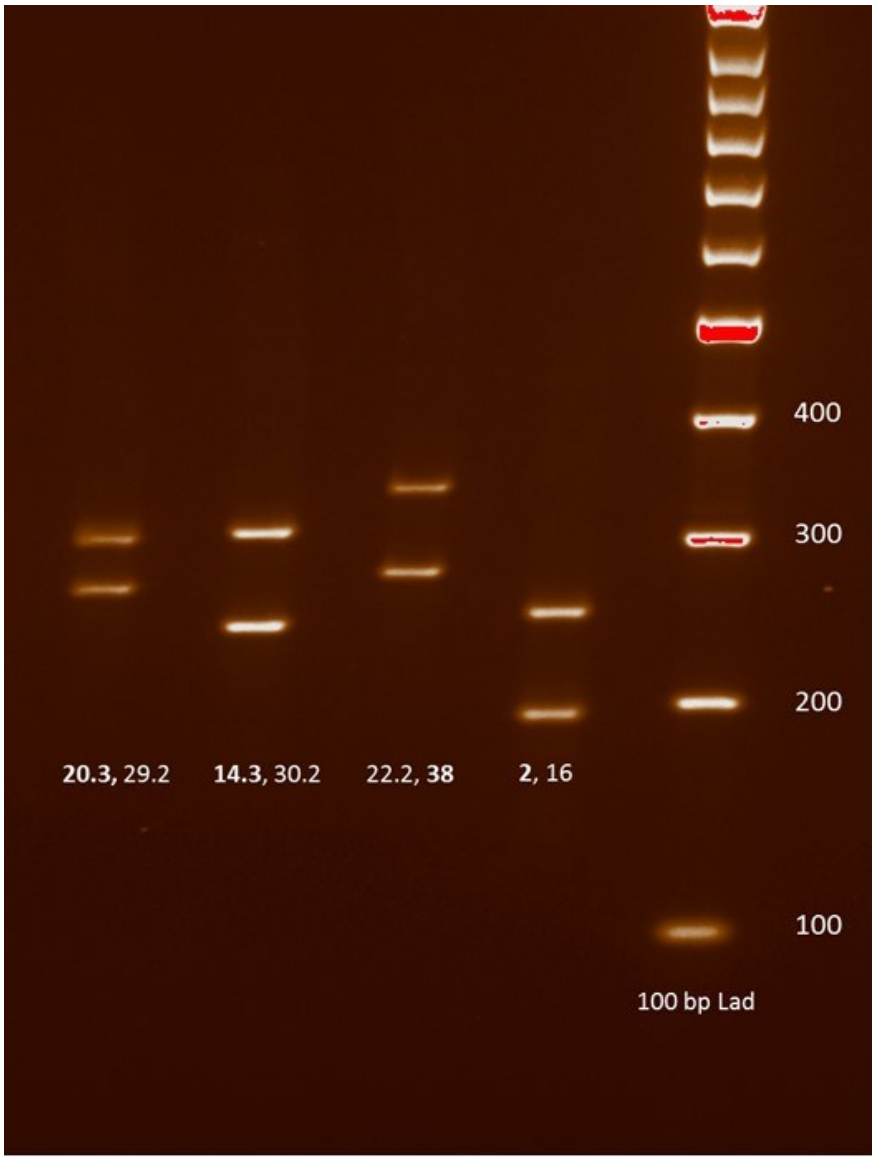
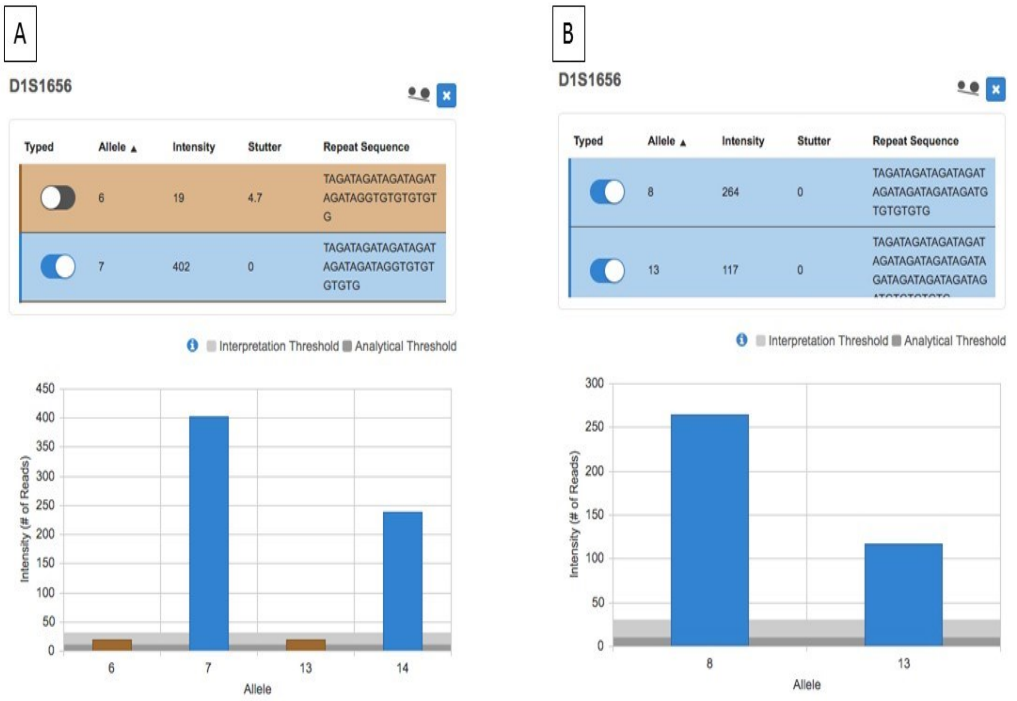


Figure 5. Sequencing data of the reverse strand of the alleles 7 and 8 at the D1S1656 locus. This data was generated using ForenSeq™ DNA Signature Prep (primer mix B) and MiSeq FGx System (Illumina, USA). (A) Shows the sequence data of allele 7; (B) Shows the sequence data of allele 8. Due to the presence of the A variant of rs78443572 SNP (TAGG, G: 73%, A: 27%) in the alleles 8 and 13, these alleles ended with TAGA rather than TAGG.




| Allele | Sequence |
|---|---|
| 7 | TAGA TAGA TAGA TAGA TAGA TAGA TAGG TGTGTGTGTG |
| 14 | TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGG TGTGTGTGTG |
| 8 | TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA TGTGTGTGTG |
| 13 | TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA TGTGTGTGTG |

**Table 1. Sequence structure of the SE33 and D1S1656 loci. (A) Shows the 5´ Local flank (15 bp), repeat region, and the 3´ Local flank (24 bp) of the SE33 locus. Type 1 alleles (< 260 bp) tend to**

**have the basic unit of [(CTTT) n] in the repeat region. Type 2 alleles (263-315 bp) have a single CTTTTT within the CTTT units. The sequence structure of larger alleles (≥ 317 bp) tend to show two CTTTTT within the CTTT units, which are called type 3 in this manuscript. It also shows an example of the sequence structure of a type 2 allele (allele 26.2, GenBank: V00481.1). (B) Shows the repeat region and the 3´ Local flank (10 bp) of the D1S1656 locus. A typical sequence structure of a D1S1656 allele is shown. The sequence structure of allele 15.3 (GenBank: G07820.1) is given for illustration. Based on the published guidelines of the International Society for Forensic Genetics (ISFG), the local flank regions showed in A and B (greyed out sequences) are not counted in allele calling system [9].**

**A. SE33 locus**

| Structure type | Size | Sequence structure: 5´ Local flank (15 bp) | Repeat region | 3´ Local flank (24 bp) |
|---|---|---|---|---|
| Type 1 | < 260 | CT $[CTTT]_2$ CCTT C | [CTTT] n | CT $[CTTT]_3$ CT $[CTTT]_2$ |
| Type 2 | 263- | CT $[CTTT]_2$ CCTT C | [CTTT] n CTTTTT [CTTT]n | CT $[CTTT]_3$ CT $[CTTT]_2$ |
| Type 3 | ≥ 317 | CT $[CTTT]_2$ CCTT C | [CTTT] n CTTTTT [CTTT] n CTTTTT | CT $[CTTT]_3$ CT $[CTTT]_2$ |
| Allele 26.2 GenBank: V00481.1 (forward strand) | | CT CTTT CTTT CCTT C CTTT CTTT CTTT CTTT CTTT CTTT CTTT CTTT CTTT CTTT CTTT CTTT CTTT CTTT CTTTTT CTTT CTTT CTTT CTTT CTTT CTTT CTTT CTTT CTTT CTTT CTTT CT CTTT CTTT CTTT CT CTTT CTTT | | |

**B. D1S1656 locus**

| Allele | Sequence structure: Repeat region | 3´ Local flank (10 bp) |
|---|---|---|
| Typical D1S1656 allele | [TAGA] n TAGG | $(TG)_5$ |
| Allele 15.3 GenBank: G07820.1 (reverse strand) | TAGA TAGA TAGA TAGA TGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGG TGTGTGTGTG | |

**Table 2. Sequence data for the forward strand of 4 previously uncharacterized SE33 alleles: 2, 14.3, 20.3, and 38. The 5´uncounted sequence (15 bp) and 3´uncounted sequence (24 bp) of the local**

**flank region; and the extended 3´ flank are shown. The amplicons sizes of the GlobalFiler kit and of primer pair (SE33-1 and SE33-2) used in this study are shown. This primer pair was used by Moller et al. [5] and Rolf et al. [6], who reported the sequence structure of the repeat region. It also shows allele names based on their sizes and based on the sequence data. Allele 2 had a 60 bp deletion in the extended 3´ flank accompanied with an additional CTTT in the 5´ Local flank (compared to the typical structure of the 5´ Local flank). Allele 14.3 had 17.1, and the 14.3 size-based classification resulted from a 14 bp deletion in the extended 3´ flank accompanied with an additional CTTT in the 5´ Local flank (compared to the typical structure of the 5´ Local flank). Allele 20.3 showed a TTT within the CTTT repeats. Although this allele has a length of 268 bp (> 263 bp), it will be classified as type 1 structure as no hexa-nucleotide repeat was observed in the repeat region. Allele 38 contained two hexanucleotide repeats within the repeat region. [(a)] Represents rs536914220 SNP at Location 6:88277143 in the alleles 2 and 14.3. [(b)] Represents rs1045867314 SNP at Location 6:88277260 in the allele 14.3. * The sequence of the 5´ Local flank observed in the alleles 20.3 and 38 is the typical sequence of this part.**

| | | | | 5´ Local flank (15 bp) | | | | | Repeat region | | | | | | | 3´ Local flank (24 bp) | | | | | Extended 3´ flank |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Allele name (size-based system) | GlobalFiler Sizes (bp) | Amplicon sizes (bp) using SE33-1 and SE33-2 | Allele name (sequence-based system) | CT | CTTT | CCTT | C | T [(a)] | CTTT | C | TTT | CTTTTT | CTTT | CTTTTT | CTTT | CT | CTTT | CT | T [(b)]T | CTTT | CTTTTTCTTTC[TTCC]$_3$TTT[CT]$_6$[CTTT]$_3$CTAA[CT]$_2$CTTTGTCT[CTTT]$_4$TGACGGAGTT |
| 2 | 296.85 | 193 | 16 | 1 | 3 | 1 | 0 | 1 | 16 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 3 | 1 | 0 | 2 | CTTTTTCTTTCTTTTTC TTC <60 bp del> [CTTT]$_2$ TGAC |

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | | | | | | | | | | GGAG TT |
| 14.3 | 348.21 | 244 | 17.1 | 1 | 3 | 1 | 0 | 1 | 2 | 1 | 0 | 0 | 15 | 0 | 0 | 1 | 3 | 0 | 1 | 2 | <14 bp del> TTC $[TTCC]_3$ TTT $[CT]_6$ $[CTTT]_3$ CTAA $[CT]_2$ CTTT GTCT $[CTTT]_4$ TGAC GGAG TT |
| 20.3 * | 372.46 | 268 | 20.3 | 1 | 2 | 1 | 1 | 0 | 9 | 0 | 1 | 0 | 11 | 0 | 0 | 1 | 3 | 1 | 0 | 2 | CTTTTT CTTT CTTTTT C $[TTCC]_3$ TTT $[CT]_6$ $[CTTT]_3$ CTAA $[CT]_2$ CTTT GTCT $[CTTT]_4$ TGAC GGAG TT |
| 38 * | 441.70 | 337 | 38 | 1 | 2 | 1 | 1 | 0 | 9 | 0 | 0 | 1 | 12 | 1 | 14 | 1 | 3 | 1 | 0 | 2 | CTTTTT CTTT CTTTTT C $[TTCC]_3$ TTT $[CT]_6$ $[CTTT]_3$ CTAA $[CT]_2$ CTTT GTCT $[CTTT]_4$ TGAC GGAG TT |