DOI:
10.1016/j.fsigss.2017.09.222

Document Version
Peer reviewed version

THE TAO OF MPS: COMMON NOVEL VARIANTS

L.A. Devesse[1], D.J. Ballard[1], L.B. Davenport[1], K.B. Gettings[2], L.A. Borsuk[2], P.M. Vallone[2], and D. Syndercombe Court[1]

[1] King's Forensics, King's College London, 150 Stamford Street, London SE1 9NH

[2] National Institute of Standards and Technology, Gaithersburg, MD, USA

Laurence.a.devesse@kcl.ac.uk

Abstract

The introduction of massively parallel sequencing (MPS) to forensic genetics has led to improvements in multiple aspects of DNA analysis, however additional complexities are concurrently associated with these advances. In relation to STR analysis, the move to assign alleles using sequence rather than length based methodologies has highlighted the extent to which previous allelic variation was masked. In this work, a series of samples (n=1000) from five different population groups (Caucasian, West African, North East African, East Asian and South Asian) were genotyped for 27 forensically validated autosomal STRs. Results were compared to data from the National Institute of Standards and Technology (NIST), with this collaborative project now providing one of the most expansive data sets generated using MPS technology to date. The large number of these variants characterised at select markers brings into question the strategies for producing representative population data, yet also provides an opportunity to utilise this diversity in unique ways. Results from this collaborative study have demonstrated that the number of samples necessary to capture the breadth of allelic variation is highly dependent on the individual marker and the extent of its sequence variability.

1. Introduction

The introduction of massively parallel sequencing (MPS) has led to an increased power of discrimination compared to traditional CE-based techniques. This is largely due to the increased number of markers that can be multiplexed, and the ability to use sequence variation to differentiate allele of the same size but differing in sequence [1-3]. In order for these "novel" sequence variants to be of use for forensic casework, new databases must be generated. Historically, 200 samples from each population group were used to capture the breadth of variation at any given autosomal STR locus, but the increased number of alleles observed using MPS brings into question whether this number is still adequate.

In this work, sequence-specific population databases were created for five UK population groups. Genotypes were obtained using the Illumina ForenSeq™ DNA Signature Prep Kit (Illumina, San Diego, CA). This report compares results from two contrasted locii with data provided by the National Institute of Standards and Technology (NIST).

2. Materials and Methods

2.1. Library preparation and sequencing

The Illumina ForenSeq™ DNA Signature Prep Kit [4] was used to prepare samples (buccal swab extracts from 1000 unrelated individuals) for sequencing on the MiSeq® FGx instrument. Primer mix A was used for the first PCR reaction, which contains primers for identity markers including 27 autosomal STRs. The only protocol modification implemented was to increase the volume of pooled libraries used for sequencing from 7 µl to 12 µl, as this has been shown to yield better results.

2.3. Data analysis and comparison

Individual STR allelic sequence variants were characterised using a modified version of STRait Razor 2.0 and in-house Excel-based workbooks [5]. Results were compared to previous results obtained by CE, and to those generated by The ForenSeq™ Universal Analysis Software (UAS) for concordance purposes.

Sequence variants observed at markers CSF1PO and D12S391 were compared to those described by NIST, and graphs were generated to show allelic diversity against number of alleles sequenced. Caucasian data from NIST was merged with White British data from King's College to generate data for the White European graph.

3. Results and discussion

Results show that the number of alleles observed at certain markers is significantly increased when using sequencing. The highest level of sequence variation was observed at D12S391, where the number of alleles in the White British population group for example increases from 18 (size based) to 53 (sequence based). As shown in Figure 1, some of the sequence variation observed was population specific, with 10 alleles at D12S391 only being seen in the White British population group.

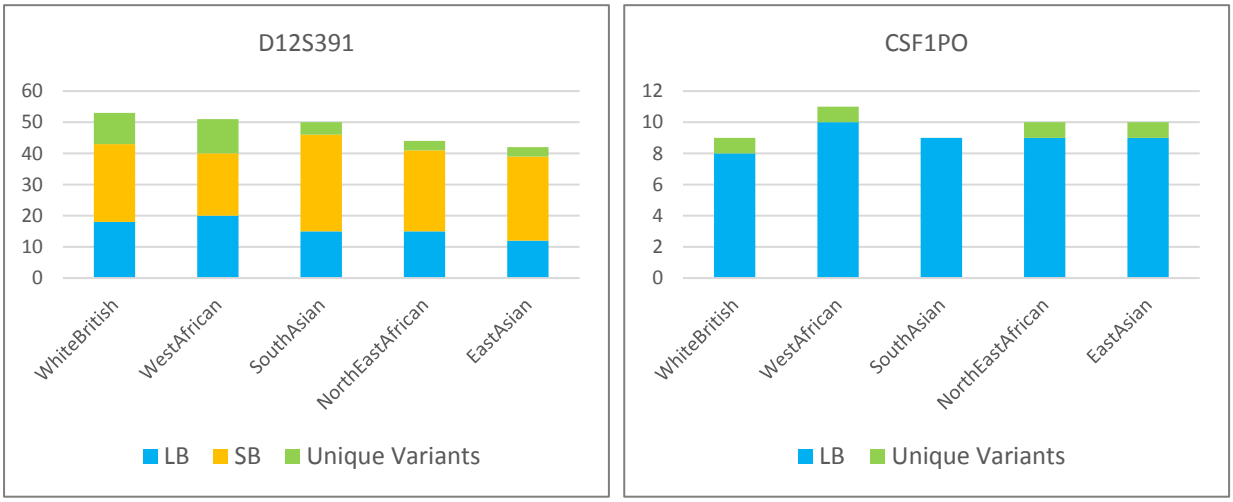

Figure 1: Increase in the number of alleles seen using sequencing at D12S391 and CSF1PO. Length based (LB) alleles are shown in blue, whilst additional sequence based (SB) alleles are shown in orange. Variants seen only in one population group are further highlighted in green (Unique Variants).

At CSF1PO, all variation observed was population specific within this dataset. In order to investigate whether these alleles are indeed population-specific, they were compared with alleles genotyped in the Caucasian, African-American and East Asian population groups by NIST and the University of North Texas (UNT) [6]. Table 1 shows that two sequence-based alleles seen only in one population group were also seen in the corresponding groups sequenced by another research group. This suggests these variants could be specific to these populations, and demonstrates the utility of larger scale databases to capture variation.

LB allele	SB allele	White British	West African	South Asian	NorthEast African	East Asian	Comments
6	[ATCT]6						
7	[ATCT]7						
8	[ATCT]8						
9	[ATCT]9						
10	[ATCT]10						
11	[ATCT]11						
	[ATCT]7ATAT[ATCT]3				1X		
12	[ATCT]12						
	[ATCT]4GTCT[ATCT]7	1X					
	[ATCT]5GTCT[ATCT]6					2X	Seen by UNT**
	[ATCT]8ACCT[ATCT]3		1X				Seen by NIST*
13	[ATCT]13						
14	[ATCT]14						
15	[ATCT]15						

Table 1: List of alleles observed at CSF1PO. Boxes coloured in blue show which population groups each allele was observed in. For sequence variants unique to one population group, the box is coloured orange and the number of times that allele was observed in the population within our data is given. One allele was also seen

during the comparison with data generated for the African American population at NIST (*), and another was seen in data published by the University of North Texas (UNT) for the East Asian Population (**) [6].

To gain an idea of how many alleles must be typed to identify the majority of sequence variants at any given STR locus, samples were randomised and alleles were plotted against novel variants observed. Figure 2 shows the resulting graph for D12S391, which shows that new variants are still being observed after 200 samples (400 alleles) typed. The addition of alleles observed within the comparable population groups at NIST show that even above 1000 alleles, i.e. 500 samples, novel alleles are still being discovered.

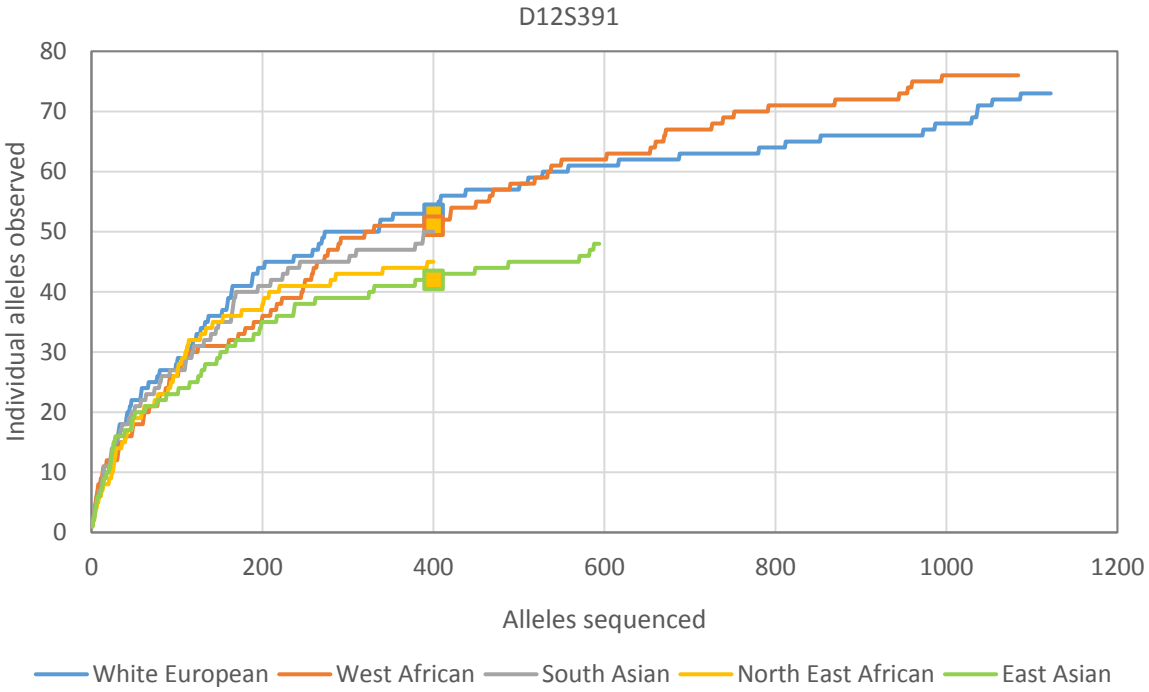

Figure 2: Graph showing the number of individual sequence-based alleles observed at D12S391 compared to the number of alleles sequenced. The yellow box on the White European, West African and East Asian graphs highlight allele 400- after which all additional alleles sequenced are from the NIST data set.

4. Conclusion

In order to capture the breadth of variation at all autosomal STR markers using massively parallel sequencing, a sample size of 200 per population group is insufficient. Larger scale studies are necessary to identify all "common sequence variants", especially at markers such as D12S391 which show a high level of variation.

References

1. Borsting, C. and N. Morling, *Next generation sequencing and its applications in forensic genetics.* Forensic Sci Int Genet, 2015. **18**: p. 78-89.
2. Gettings, K.B., K.M. Kiesler, S.A. Faith, E. Montano, C.H. Baker, B.A. Young, R.A. Guerrieri, and P.M. Vallone, *Sequence variation of 22 autosomal STR loci detected by next generation sequencing.* Forensic Sci Int Genet, 2016. **21**: p. 15-21.
3. Churchill, J.D., S.E. Schmedes, J.L. King, and B. Budowle, *Evaluation of the Illumina((R)) Beta Version ForenSeq DNA Signature Prep Kit for use in genetic profiling.* Forensic Sci Int Genet, 2016. **20**: p. 20-9.
4. Illumina, *ForenSeq™ DNA Signature Prep Reference Guide.* Document #15049528 v01, 2015.
5. Warshauer, D.H., J.L. King, and B. Budowle, *STRait Razor v2.0: The improved STR Allele Identification Tool – Razor.* Forensic Science International: Genetics, 2015. **14**(Supplement C): p. 182-186.
6. Novroski, N.M., J.L. King, J.D. Churchill, L.H. Seah, and B. Budowle, Characterization of genetic sequence variation of 58 STR loci in four major population groups. Forensic Sci Int Genet, 2016. 25: p. 214-226.