Forensic Science International: Genetics Supplement Series 5 (2015) e121–e123

Contents lists available at ScienceDirect

Forensic Science International: Genetics Supplement Series

journal homepage: www.elsevier.com/locate/FSIGSS

The next dimension in STR sequencing: Polymorphisms in flanking regions and their allelic associations

Katherine Butler Gettings[a,*], Rachel A. Aponte[b], Kevin M. Kiesler[a], Peter M. Vallone[a]

[a] U.S National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, MD 20899-8314, USA
[b] Department of Forensic Sciences, The George Washington University, Washington, DC 20007-1150, USA

ARTICLE INFO

Article history:
Received 26 August 2015
Accepted 14 September 2015
Available online 24 September 2015

Keywords:
STR loci
SNP
Next generation sequencing

ABSTRACT

Sequence data was analyzed from the flanking and repeat regions of 22 autosomal STR loci in 183 population samples. The flanking regions of the loci sequenced in this data set can be divided into six categories, based on the types of polymorphisms they contain: (1) multiple polymorphisms in haplotype, (2) "old" single polymorphisms, (3) population or allele specific polymorphisms, (4) polymorphisms associated with sequence variants, (5) rare polymorphisms, and (6) no polymorphisms. These results provide an indication of which loci are expected to gain in areas such as allelic diversity and ancestry information by including flanking regions in STR sequence analysis.

Published by Elsevier Ireland Ltd.

1. Introduction

Recent publications have shown potential gains from sequencing the repeat regions of forensic STR loci [1,2]; however, the potentially polymorphic flanking regions have not previously been analyzed in a comprehensive manner. Including flanking regions in the analyzed sequence data will add to allelic diversity, could aid kinship interpretation, and may improve our understanding of mutational events and evolutionary history at these loci. Due to their increased stability/lower mutation rate, the flanking region polymorphisms may also help inform nomenclature decisions. On a more practical level, knowledge of these flanking region polymorphisms allows for improved amplification primer design, bioinformatic search algorithm development, and configuration of bioinformatic pipelines in a manner which maintains back-compatability to existing STR genotyping results. Publicly available data can provide insight into the global allele frequency of polymorphisms close to STR repeat regions [3], but due to alignment difficulties, these data sets do not currently include repeat region sequences. Therefore, studies such as this are needed to quantify the expected gains in allelic diversity and ancestry information by including flanking regions in STR sequence analysis.

2. Materials and methods

Generation of the sequence data analyzed herein, as well as repeat region variation, has been previously described [2]. In this current study, sequence data from 22 autosomal STR loci in 183 population samples (70 European, 68 African American, and 45 Hispanic individuals) was further analyzed using the STRait Razor (version 1.5) algorithm [4] with a custom configuration file encompassing the entire amplified and sequenced region. The resulting sequences were aligned within each locus using Geneious (version 8.1.4), and flanking region polymorphisms were cataloged and cross-referenced to repeat region alleles and populations.

3. Results and discussion

The flanking regions of the loci sequenced in this data set can be divided into six categories, based on the types of polymorphisms they contain: (1) multiple polymorphisms in haplotype, (2) "old" single polymorphisms, (3) population or allele specific polymorphisms, (4) polymorphisms associated with sequence variants, (5) rare polymorphisms, and (6) no polymorphisms. These categories are described in reverse order (from least to most gain) below, along with the loci that fall into each category and examples of expected gains. This is not a comprehensive list; sequencing a greater extent of flanking region at these loci, sequencing additional loci, and sequencing additional samples may reveal more categories.

* Corresponding author.
E-mail address: katherine.gettings@nist.gov (K.B. Gettings).

http://dx.doi.org/10.1016/j.fsigss.2015.09.049
1875-1768/Published by Elsevier Ireland Ltd.

3.1. No polymorphisms

Within this 183 sample set and for the extent to which the flanking regions were sequenced (ranging from 11 bp to 167 bp), six loci showed no flanking region variation: D3S1358, FGA, D12S391, Penta E, D19S433, and D21S11.

3.2. Rare polymorphisms

Similar to rare repeat region sequence variants, single flanking region variants observed at approximately <5% frequency may be helpful for a particular sample or case but will not contribute greatly to the allelic diversity expected at a locus. Within this 183 sample set, seven loci show examples of rare flanking region variants: D2S441, CSF1PO, D8S1179, D10S1248, D18S51, Penta D, and D22S1045. Sequencing more individuals at these loci may allow some of these SNPs to be associated to particular STR alleles and/or populations.

3.3. Polymorphisms associated with a sequence variant

Polymorphisms in this category are not expected to add to allelic diversity, as they appear to be linked to a particular repeat region sequence variant. Examples of this are found at the D1S1656 and vWA loci. At the D1S1656 locus, the minor allele T of SNP rs4847015 appears to be associated with .3 STR alleles. The .3 STR motif is caused by the presence of a trinucleotide following 3 to 4 tetranucleotide repeats: $[TAGA]_3$ TGA $[TAGA]_{11}$ [TAGG] = 15.3. In this sample set, these microvariants were present in sizes ranging from 15.3 to 19.3 and totaled over 26% of D1S1656 alleles. In every instance of a .3 STR allele across all populations in this data set, the minor T allele is present at rs4847015, and for all other STR alleles the C allele is present. Despite the widespread distribution across populations and high frequency of this SNP, it is not expected to improve discrimination at this locus.

3.4. Population or allele specific polymorphisms

For some loci, marginal gains in allelic diversity are possible due to flanking region SNPs that appear to be associated with one population and often with a particular repeat region allele. This pattern of distribution suggests a more recent occurrence, subsequent to human population divergence and without sufficient time for the repeat region of alleles containing the flanking region SNP to expand/contract, which would more broadly distribute the SNP across allele sizes. Examples of this can be observed at the TPOX and TH01 loci. At the TPOX locus, SNP rs13422969 appears to be primarily associated with the 9 STR allele in individuals of African origin. Within the 136 African American chromosomes genotyped in this data set, the minor A allele was observed with a frequency of 13.5% overall and a frequency of 64% for the 9 STR allele. It was observed once with a 10 STR allele in the African American samples, and once with a 9 STR allele among the 90 Hispanic chromosomes genotyped. While this SNP will not offer a large improvement in discrimination at this locus, the presence of an A allele could contribute to an ancestry prediction.

3.5. "Old" single polymorphisms

When one SNP which is well distributed across populations and allele sizes/sequences is found in the flanking region, an increase in alleles will routinely be observed. The flanking region of D16S 539 contains an example of this with SNP rs1728369, where the minor allele C of this SNP is well distributed across the most common STR alleles 11, 12, and 13 (in 183 population samples), in addition to being distributed across populations (both in the 183 population samples and in 1000 Genomes data).

3.6. Multiple polymorphisms in haplotype

The most gain from sequencing and analyzing the flanking region occurs in loci which contain multiple flanking region SNPs, where the resulting haplotypes are well-distributed across populations and allele sizes/sequences. Three simple repeat loci show this pattern within this data set: D5S818, D7S820 and D13S317. These loci do not gain significant numbers of alleles by sequencing the repeat region, but analyzing the flanking region results in 2 to 3 times more alleles. D5S818 has a four SNP haplotype composed of rs25768, rs146841551, rs541272009, and a G/T SNP 4 bp downstream of the repeat region. D7S820 contains a three SNP haplotype including rs16887642, rs7789995, and rs7786079. D13S317 has a four SNP plus two InDel haplotype: rs9546005, rs73525369, rs735250432, rs146621667, a 4 bp deletion 8 bp downstream, and a 4 bp deletion 21 bp downstream (InDel placement may vary based on alignment parameters). Fig. 1 shows the diversity of alleles obtained via analysis of the flanking regions for each of these loci. Linkage disequilibrium (LD) analysis was performed using Arlequin (version 3.5.2.2) for polymorphisms found at >5% frequency at these loci. The highest r^2 value observed for these SNP pairs was 0.1586. Although this may provide evidence of LD, which should limit the use of these markers as independent loci, this is not a factor when using these polymorphisms in haplotype with the STR allele.

4. Conclusions

Ideal polymorphisms for increasing allelic diversity are neither population nor allele specific. In addition, multiple SNPs in haplotype may substantially increase the diversity of alleles,

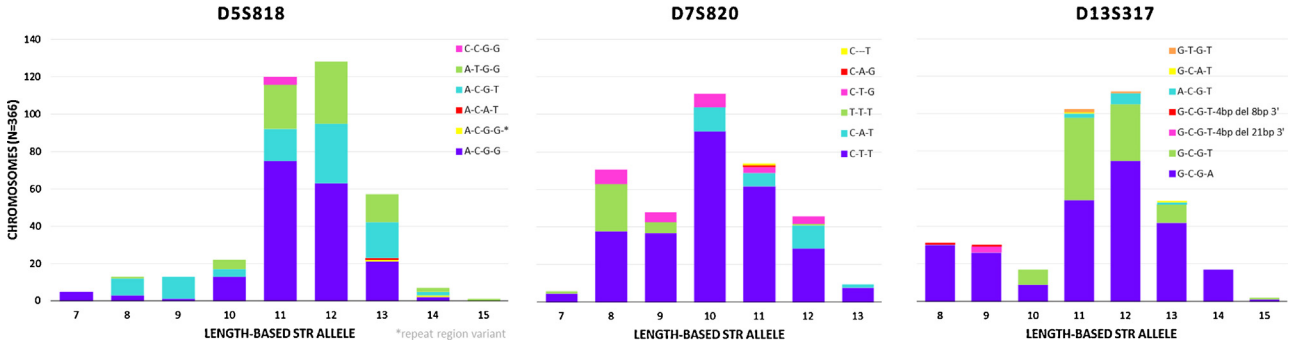

Fig. 1. Sequence haplotypes observed from flanking region polymorphisms for each length-based allele at D5S818, D7S820, and D13S317. Data from 183 individuals (366 chromosomes) in three populations.

particularly at simple repeat loci. Lastly, as the publicly available data, such as that available through the 1000 Genomes project, does not contain repeat region sequences, population studies are needed to associate flanking region polymorphisms with STR alleles in order to generate haplotype frequency data for forensic casework statistics.

Role of funding

This work was funded in part by the Federal Bureau of Investigation (FBI) interagency agreement DJF-13-0100-PR-0000080: "DNA as a Biometric". Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

Conflict of interest

None.

Acknowledgements

None.

References

[1] C. Gelardi, E. Rockenbauer, S. Dalsgaard, C. Borsting, N. Morling, Second generation sequencing of three STRs D3S1358, D12S391 and D21S11 in Danes and a new nomenclature for sequenced STR alleles, Forensic Sci. Int. Genet. 12 (2014) 38–41.

[2] K.B. Gettings, K.M., Kiesler, S.A., Faith, R.A., Guerrieri, B.A., Young, C.H. Baker, et al. Sequence variation of 22 autosomal STR loci detected by next generation sequencing, Manuscript submitted 2015.

[3] The 1000 Genomes Project Consortium, An integrated map of genetic variation from 1092 human genomes, Nature 491 (2012) 56–65.

[4] D.H. Warshauer, D. Lin, K. Hari, R. Jain, C. Davis, B. Larue, et al., 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.