Genomics Proteomics Bioinformatics 13 (2015) 250–257

HOSTED BY

ELSEVIER

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

ORIGINAL RESEARCH

Novel Y-chromosome Short Tandem Repeat Variants Detected Through the Use of Massively Parallel Sequencing

David H. Warshauer [1,a], Jennifer D. Churchill [1,b], Nicole Novroski [1,c], Jonathan L. King [1,d], Bruce Budowle [1,2,*,e]

[1] *Institute of Applied Genetics, Department of Molecular and Medical Genetics, University of North Texas Health Science Center, Fort Worth, TX 76107, USA*
[2] *Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah 21589, Saudi Arabia*

Received 25 May 2015; revised 28 July 2015; accepted 19 August 2015
Available online 21 September 2015

Handled by Changqing Zeng

KEYWORDS

Y-STR;
Sequence polymorphism;
Allele variants;
Massively parallel sequencing;
Nextera;
STRait Razor

Abstract **Massively parallel sequencing** (MPS) technology is capable of determining the sizes of short tandem repeat (STR) alleles as well as their individual nucleotide sequences. Thus, single nucleotide polymorphisms (SNPs) within the repeat regions of STRs and variations in the pattern of repeat units in a given repeat motif can be used to differentiate alleles of the same length. In this study, MPS was used to sequence 28 forensically-relevant Y-chromosome STRs in a set of 41 DNA samples from the 3 major U.S. population groups (African Americans, Caucasians, and Hispanics). The resulting sequence data, which were analyzed with **STRait Razor** v2.0, revealed 37 unique allele sequence variants that have not been previously reported. Of these, 19 sequences were variations of documented sequences resulting from the presence of intra-repeat SNPs or alternative repeat unit patterns. Despite a limited sampling, two of the most frequently-observed variants were found only in African American samples. The remaining 18 variants represented allele sequences for which there were no published data with which to compare. These findings illustrate the great potential of MPS with regard to increasing the resolving power of STR typing and emphasize the need for sample population characterization of STR alleles.

* Corresponding author.
E-mail: Bruce.Budowle@unthsc.edu (Budowle B).
[a] ORCID: 0000-0002-6653-5644.
[b] ORCID: 0000-0001-9796-193X.
[c] ORCID: 0000-0001-9071-9278.
[d] ORCID: 0000-0003-4541-1789.
[e] ORCID: 0000-0003-4116-2930.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

http://dx.doi.org/10.1016/j.gpb.2015.08.001

Introduction

Short tandem repeat (STR) markers located on the Y-chromosome (Y-STRs) are extremely useful because of a lack of recombination. Barring mutation, all paternally-related males share the same Y-STR haplotype. As a result, Y-STRs are used in genealogical and evolutionary studies, and forensic genetics casework including paternity testing to determine the biological father of a particular male child, missing persons cases where the Y-STR haplotype can serve as an extended reference profile for a given paternal lineage, and analyses of mixture evidence where there is substantially more female DNA than male DNA. Indeed, the variety of uses for Y-STR markers has made them the object of extensive research and application within the scientific community.

Given the value of STR markers in identity testing, efforts are underway to increase the power of discrimination associated with their respective typing and analysis methods. Primarily, an increase in power of discrimination has been accomplished through the introduction of new, highly-polymorphic STRs and by developing larger multiplex panels [1–4]. Discrimination power also may be increased by further characterization beyond nominal length of the alleles at extant loci. STR alleles are typically characterized by the number of units in their repeat motifs, a distinction commonly determined by size separation by capillary electrophoresis (CE). However, other detection methods, such as Sanger sequencing and mass spectrometry, have been used to determine both the size and the nucleotide composition of STR alleles [5,6]. The emergence of massively parallel sequencing (MPS) technologies improved upon this principle by allowing for the detection of a substantially-larger amount of genetic sequence information with a higher throughput, lower cost, and greater ease-of-use than previous methods. Studies involving each of these approaches have resulted in the detection of intra-repeat single nucleotide polymorphisms (SNPs) and novel repeat motif variants, which allow for a greater level of distinction than that of traditional CE methods [5–10]. For instance, two individuals with the same nominal allele(s) (based on length) at a certain locus potentially may be distinguished by MPS if the nucleotide sequence of the allele differs between them. This level of resolution may prove invaluable in the deconvolution of genetic mixtures and also could provide information about population-specific alleles for evolutionary studies.

In this proof-of-principle study, MPS was used to determine the repeat sequences of 28 forensically-relevant Y-STRs across a dataset of three major US populations (n = 41): Caucasians (CAU), Hispanics (HIS), and African Americans (AFA). These sequence data revealed several intra-repeat SNPs and allelic variants that have not been documented previously. The novel variants described herein are indicative of the potential of MPS with regard to identifying additional genetic diversity of Y-STRs and support that more in depth population studies are warranted.

Results

Since nanogram and subnanogram quantities of input DNA can be typed by MPS, PCR enrichment has become the method of choice for studies involving forensic applications. However, this study employed a capture enrichment approach. The TruSeq library preparation chemistry was selected initially, because no PCR amplification is required. Therefore, primer binding site mismatch issues would not impact multiplex design or the amplification success. It was hypothesized that a dense probe design would increase capture efficiency of the target loci. In addition, PCR-generated errors would be reduced, thus minimizing potential artifacts. Lastly, laying a foundation of sequence data with an alternate enrichment system could be useful when full validation studies are undertaken.

Sequencing coverage

All 28 Y-STR loci were detected with the approach described herein. The coverage ranged from 0 to 1493 ×, with a mean coverage of 9–387 ×. The lowest performing markers were DYS448 (mean 9 ×), DYS449 (mean 33 ×), DYS518 (mean 34 ×), DYS389II (mean 37 ×), and DYS505 (mean 38 ×); while the highest were DYS643 (mean 322 ×), DYS391 (mean 333 ×), and DYS438 (mean 387 ×).

Sequence variants

A total of 37 unique Y-STR allele sequences that have not been previously published were detected across the 41 samples used in this study. These sequences may be divided into 2 categories: nominal allele variant sequences and novel allele sequences. For the purposes of this study, a nominal allele variant sequence is defined as any allele sequence that differs from the previously-documented sequence(s) for that particular allele, whereas a novel allele sequence refers to the sequence detected for an allele that has no previously published sequence data with which to compare.

Nominal allele variants

Of the 37 previously-undocumented allele sequences that were detected, 19 were classified as nominal allele variant sequences. These nominal variants were found in loci DYS389I/II, DYS390, DYS393, DYS481, DYS518, and DYS635, and have been further characterized as either SNP variants or repeat pattern variants (RPVs) (**Table 1**). Allele sequence variation may be introduced via strand slippage or one or more point variations within the repeat region. In this study, nominal variant sequences were classified as SNP variants if they displayed a repeat motif that differs from the commonly-described motif, an occurrence indicative of point substitution. RPVs are defined as allele sequences that differ from published data with regard to repeat unit arrangement, but are consistent with the reported repeat motif. Such variations may be due to strand slippage or the presence of intra-repeat SNPs, but definitive conclusions cannot be made without additional data. To illustrate the differences between these two types of variants, consider a locus with a reported repeat motif of $[TCTA]_n[TCTG]_p$ (where n and p represent the number of repeats). If a "17" allele was detected with a repeat motif of $[TCTA]_5$**[TATA]$_1$** $[TCTG]_{11}$, this nominal allele variant sequence would likely be due to the presence of a C/A SNP in the first "TCTA" repeat unit. Since such a change results in a "TATA" repeat unit that is inconsistent with the reported repeat motif, this sequence would be classified as a SNP variant. However, if another nominal variant was detected for this allele with a repeat motif of $[TCTA]_6[TCTG]_{11}$, it would be labeled a