Forensic Science International: Genetics Supplement Series 7 (2019) 218–220

Contents lists available at ScienceDirect

Forensic Science International: Genetics Supplement Series

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

Unleashing novel STRS via characterization of genome in a bottle reference samples

Katherine B. Gettings, Lisa A. Borsuk, Justin Zook, Peter M. Vallone

NIST, 100 Bureau Drive, 20899, Gaithersburg, MD, United States

ARTICLE INFO

Keywords:
STR
Sequence
Reference material

ABSTRACT

The higher level of multiplexing possible with current sequencing technologies encourages adoption of additional STR loci to aid in mixture interpretation [1]. However, characterization of these loci and orientation on the human genome is vital for interlaboratory comparability and databasing. Currently, when a laboratory publishes population data from a locus not previously characterized for forensic use, there is no robust way to verify the locus designation, repeat region format, and fidelity of target. To address this, we have evaluated short- and long-read sequence data generated for reference materials included in the Genome in a Bottle Consortium (GIAB) [2] with the goal of reporting STR sequences for loci which may be of interest to the forensic community. Initially, we have analyzed GIAB data using Marshfield sets of primers (published in [3]), targeting over 600 microsatellite loci with STRaitRazor 3.0 [4]. In the future, this approach can be expanded to include other loci of interest. High-confidence STR sequence data will be made publicly available via GenBank record creation within the STRSeq BioProject [5]. As the cell lines represented in GIAB reference materials are available for purchase, this STR dataset represents a robust method for researchers to confirm targeted loci.

1. Introduction

Genome in a Bottle (GIAB) is a public-private-academic consortium hosted by NIST which provides authoritative characterization of human genomes for use in clinical analytical validation and technology development. The seven GIAB samples are sequenced to varying degrees with Illumina HiSeq (PCR-free library preparation), PacBio, Oxford Nanopore, and 10X Genomics technologies. Sequence Data and VCF files are available for GRCh37 and GRCh38 under each genome at ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp. The work described herein explores the possibility of using GIAB data to support forensic STR sequencing initiatives.

2. Methods

In this proof of concept study, 669 autosomal STR targets were identified and these regions of PacBio and Illumina sequencing data were extracted from one GIAB sample, HG002. Custom STRaitRazor 3.0 configuration files were designed for the two data types: Illumina analysis was configured with 10 bp recognition sites adjacent to the repeat and PacBio analysis was configured using published amplification primer sequences as recognition sites. Post-processing, STRaitRazor outputs were triaged by 1) targets returning sequences of the expected repeat motif in both platforms, 2) targets containing the expected repeat motif with clear results in only one platform, 3) targets that failed to return the expected sequence/motif. For results in category 1, average read depth, forward/reverse (F/R) balance (forward strand read depth divided by total read depth), and allele coverage ratio (ACR, lower read depth value divided by higher read depth value in heterozygous pairs) were calculated. Troubleshooting was performed using Integrated Genomics Viewer (IGV).

3. Results and discussion

Of the 669 autosomal STR loci targeted, 377 loci (59%) returned sequences of the expected motif from both Illumina and PacBio data. On average, read depths were 79X for Illumina and 40X for PacBio. Forward/Reverse (F/R) Balance and Allele Coverage Ratios (ACR) were comparable across platforms, with F/R balance averaging 0.50 (PacBio) to 0.51 (Illumina), and ACR averaging 0.83 for both platforms (see Fig. 1 for comparison of metrics across platforms). Homozygous alleles account for 27% of the successful targets; the high average heterozygote ACR lends confidence to these homozygous calls. Instances of a locus appearing homozygous in one platform and heterozygous in the other were investigated and explained by the differing coverage ranges (see IGV display at the ATA44G07 M locus, Fig. 2). Additionally,

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

https://doi.org/10.1016/j.fsigss.2019.09.084
Received 5 September 2019; Accepted 25 September 2019
Available online 26 September 2019
1875-1768/ Published by Elsevier B.V.

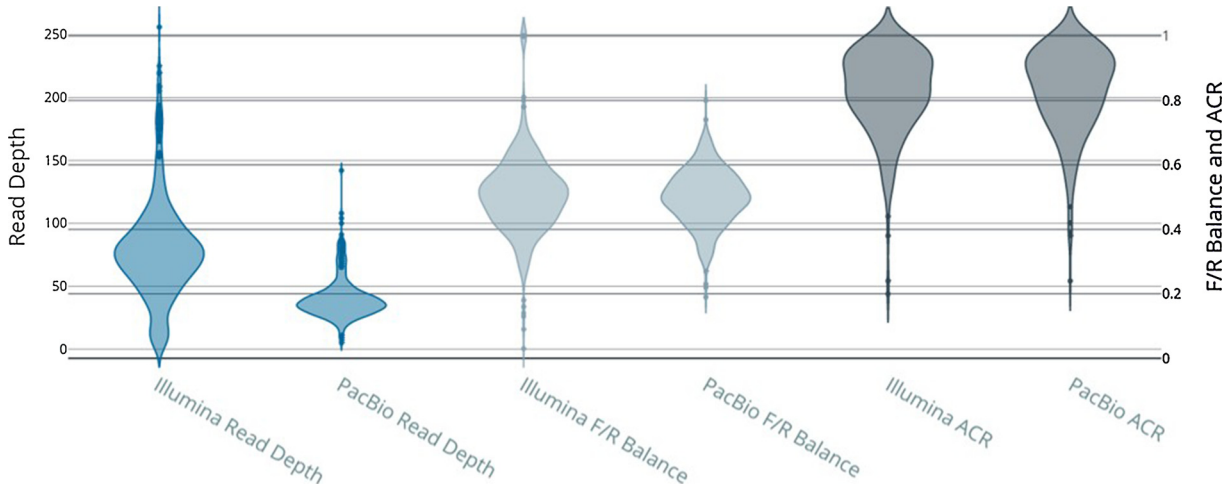

Fig. 1. Read depth, Forward/Reverse (F/R) balance, and Allele Coverage Ratios (ACR) for Illumina and PacBio GIAB data in 377 STR regions.

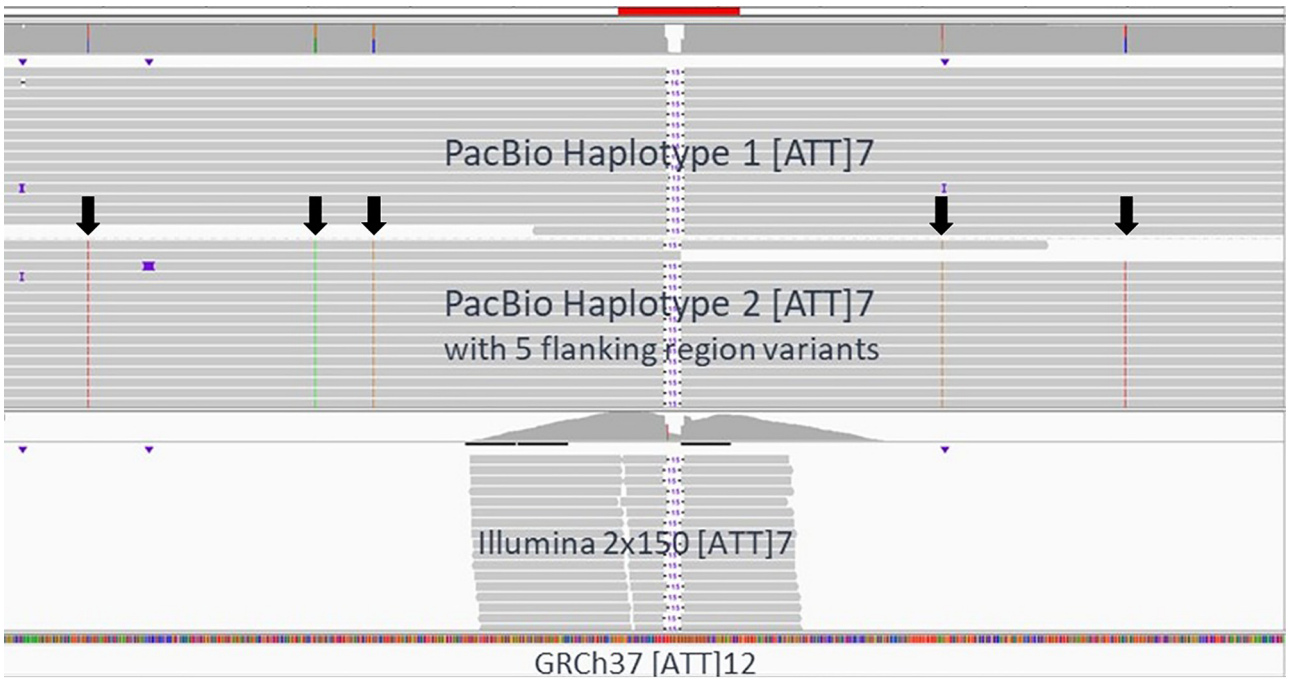

Fig. 2. IGV display of results at the ATA44G07M locus, showing five flanking region polymorphisms (black arrows) in one of the PacBio haplotypes which are outside of the targeted Illumina data range.

concordant results were obtained for 23 commonly used forensic STR loci when comparing the genotypes generated through this analysis and previous genotype data from forensic kit-based sequencing data.

For the remaining 41% of loci which did not return expected sequences in both data sets, approximately 30% clearly contain the expected motif in the PacBio data but not in the Illumina data, and 2% were present in Illumina but not PacBio. It is expected that a redesign of the applicable recognition sites would greatly improve the overall success. Finally, approximately 9% of the loci targeted did not return results consistent with the expected sequence/motif for either data set and would require additional troubleshooting.

4. Conclusions

When STR loci of interest are identified, the sequences extracted from GIAB genome data can be cataloged in the STRSeq BioProject [5] at NCBI. This will make the information readily available to the forensic community. Below we consider three possible use cases for this information:

4.1. STR Marker Discovery/Evaluation

Researchers considering "novel" STR loci for forensic use might begin with resources such as the STR Catalog Viewer (strcat.teamerlich.org, [6]). Once targets have been identified, the high quality/long read data from GIAB genomes may serve as an additional evaluation tool and aid in assay design.

4.2. QC for Novel STR Assay Targets

GIAB sequences and associated Coriell or NIST RM samples could serve as a positive control for novel STR targets.

4.3. Input for Nomenclature Discussions

High quality and longer read GIAB sequences can serve as

additional exemplar data for STR sequence nomenclature decisions.

We are interested in feedback from the forensic community regarding other uses of this resource. Please contact strseq@nist.gov to continue the conversation.

Role of funding

This work was funded by the NIST Special Programs Office and the FBI Biometric Center of Excellence Unit: DNA as a Biometric.

Declaration of Competing Interest

None.

Acknowledgements

Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Commerce. Certain commercial equipment, instruments, and materials are identified in order to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or endorsement by NIST, nor does it imply that any of the materials, instruments, or equipment identified are necessarily the best available for the purpose. This work was approved by the NIST Human Subjects Protection Office.

References

[1] N.M.M. Novroski, et al., Expanding beyond the current core STR loci: an exploration of 73 STR markers with increased diversity for enhanced DNA mixture deconvolution, Forensic Sci. Int. Genet. 38 (January) (2019) 121–129.

[2] J. Zook, et al., Reproducible integration of multiple sequencing datasets to form high-confidence SNP, indel, and reference calls for five human genome reference materials, bioRxiv (2018) 281006, https://doi.org/10.1101/281006.

[3] T.J. Pemberton, et al., Sequence determinants of human microsatellite variability, BMC Genomics 16 (10) (2009 Dec) 612.

[4] A.E. Woerner, et al., Fast STR allele identification with STRait razor 3.0, Forensic Sci. Int. Genet. 30 (September) (2017) 18–23.

[5] K.B. Gettings, et al., STRSeq: a catalog of sequence diversity at human identification Short Tandem Repeat loci, Forensic Sci. Int. Genet. 31 (November) (2017) 111–117.

[6] T. Willems, et al., The landscape of human STR variation, Genome Res. 24 (November) (2014) 1894–1904.