Unleashing Novel STRs

via characterization of

Genome in a Bottle

reference samples

Katherine Butler Gettings[1]
Lisa A. Borsuk[1]
Justin Zook[2]
Peter M. Vallone[1]

What is GIAB?

A consortium hosted by NIST dedicated to

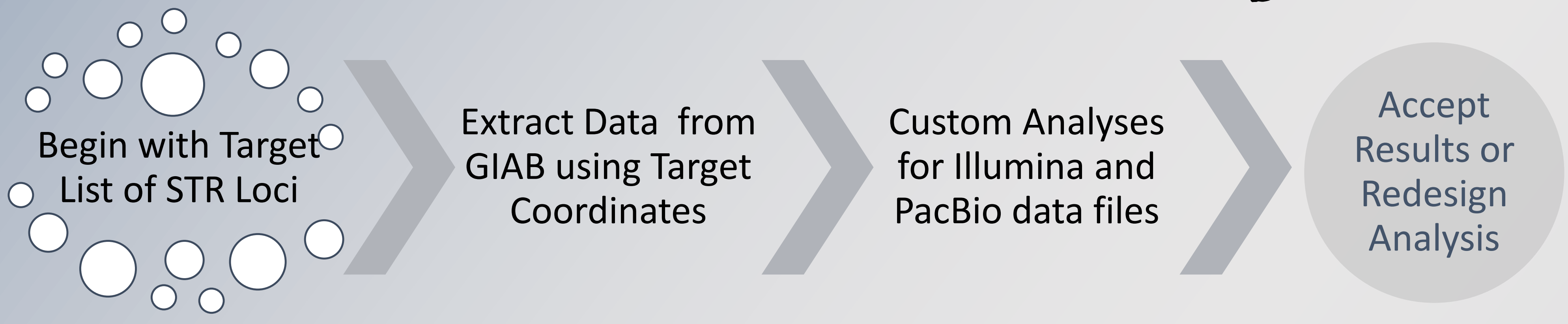
AUTHORITATIVE CHARACTERIZATION

of benchmark human genomes.

Genome	Coriell cell line ID	NIST ID	NIST RM #	NCBI BioSample	PGP ID
CEPH Mother/Daughter	GM12878	HG001	RM8398	SAMN03492678	Not PGP
AJ Son	GM24385	HG002	RM8391(son)/ RM8392(trio)	SAMN03283347	huAA53E0
AJ Father	GM24149	HG003	RM8392(trio)	SAMN03283345	hu6E4515
AJ Mother	GM24143	HG004	RM8392(trio)	SAMN03283346	hu8E87A9
Chinese Son	GM24631	HG005	RM8393	SAMN03283350	hu91BD69
Chinese Father	GM24694	N/A	N/A	SAMN03283348	huCA017E
Chinese Mother	GM24695	N/A	N/A	SAMN03283349	hu38168C

All seven genomes are available as Coriell cell lines, while five are also available as NIST reference materials (RM). Six of the cell lines derive from Personal Genome Project (PGP) samples.

How do we extract STR data from GIAB genomes?



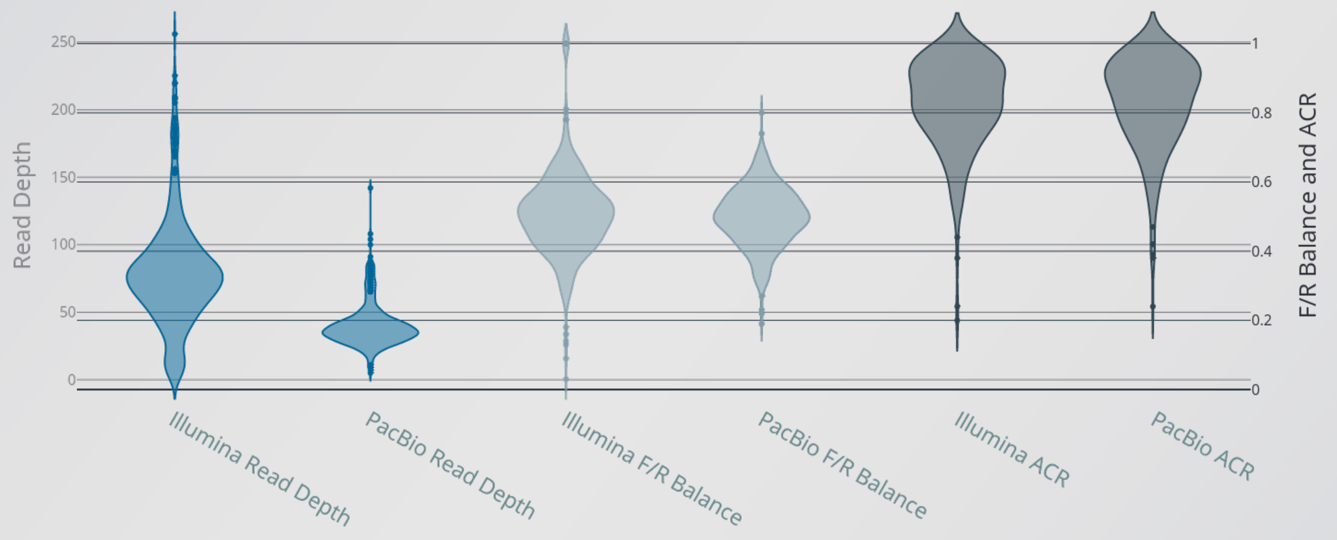
Begin with Target List of STR Loci

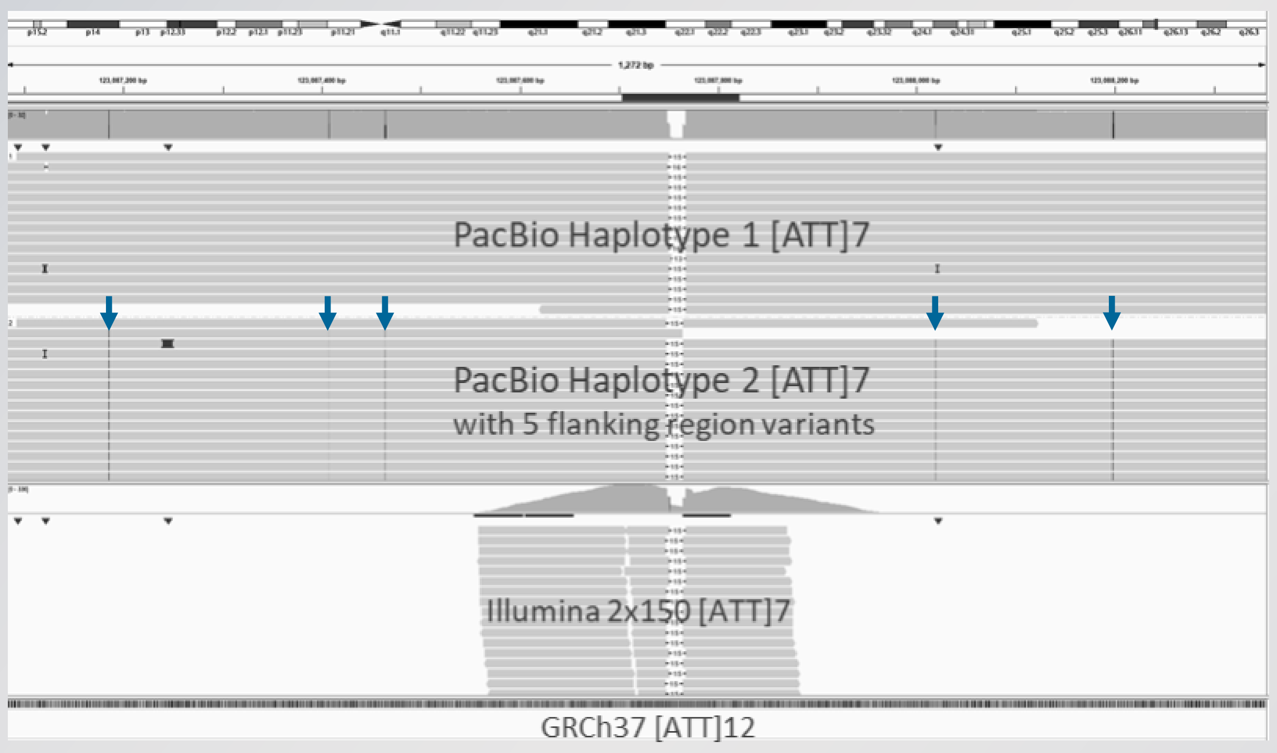
Extract Data from GIAB using Target Coordinates

Custom Analyses for Illumina and PacBio data files

Accept Results or Redesign Analysis

In a proof of concept analysis of 669 targets in one GIAB sample, **377 targets returned sequences from both Illumina and PacBio.** 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.




PacBio Haplotype 1 [ATT]7
PacBio Haplotype 2 [ATT]7
with 5 flanking region variants
Illumina 2x150 [ATT]7
GRCh37 [ATT]12

Manual inspection of results in IGV at the ATA44G07M locus, where the returned sequences indicate an isoallele in the PacBio analysis range but not in the Illumina range. PacBio sequences can be separated by haplotype to reveal one allele has 5 SNPs (blue arrows), which are beyond the bounds of the extracted Illumina data.

How can we use this resource?

STR Marker DISCOVERY/ Evaluation

QC for novel STR ASSAY targets

Input for NOMENCLATURE discussions

ISFG P636

[1]National Institute of Standards and Technology, Applied Genetics Group, Gaithersburg, Maryland, USA.
[2] National Institute of Standards and Technology, Bioassay Methods Group, Gaithersburg, Maryland, USA.

[1] Novroski NMM, et al. Forensic Sci Int Genet. 2019 Jan; 38:121-129.
[2] Zook J, et al. bioRxiv 281006; doi: https://doi.org/10.1101/281006.
[3] Pemberton TJ, et al. BMC Genomics. 2009 Dec 16; 10:612.
[4] Woerner AE, et al. Forensic Sci Int Genet. 2017 Sep; 30:18-23.
[5] Gettings KB, et al. Forensic Sci Int Genet. 2017 Nov; 31:111-117.
[6] Willems T, et al. Genome Res. 2014 Nov; 24: 1894-1904.

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.

METHODS

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. GIAB samples (see table, left) 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.

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).

RESULTS

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. 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 ATA44G07M locus, left). Additionally, 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.

USE CASES

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:

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.

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.

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.

Fine Print: "flask" is used under CC0 1.0 **Disclaimer:** 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. **Funding:** NIST Special Programs Office: Forensic DNA and the FBI Biometric Center of Excellence Unit: DNA as a Biometric.