Unleashing Novel STRs via characterization of Genome in a Bottle reference samples

What is GIAB?
A consortium hosted by NIST dedicated to AUTHORITATIVE CHARACTERIZATION of benchmark human genomes.

How do we extract STR data from GIAB genomes?

How can we use this resource?

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 genome 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-trac.ebi.ac.uk/ceg/nih/.

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 STnRaitRazor 3.0 configuration files were designed for the two data sets. Illumina analysis was performed with the 10 bp recognition sites adjacent to the repeat and PacBio analysis was configured using published amplification primer sequences as recognition sites. Post-processing, STnRaitRazor outputs were triaged by 1) targets returning sequences of the expected repeat motif in both platforms, 2) targets returning clean 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 ratios (ACR) lower than the value divided by the expected repeat motif could be estimated. ATRaitRazor outputs were triaged by 1) targets returning sequences of the expected repeat motif for either data set, 2) targets returning clean results in only one platform, 3) targets that failed to return the expected sequence/motif.

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 77X 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 differing coverage ranges (see IGV display at the ATAA4G07M locus, left). Additionally, concomitant results were obtained for 91 concomitant use, forensic STR loci when comparing the Genome in a Bottle and the Genome in a Bottle Consortium. This approach can be expanded to include other loci of 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-trac.ebi.ac.uk/ceg/nih/.

In the future, this approach can be expanded to include other loci of 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-trac.ebi.ac.uk/ceg/nih/.

DISCLAIMER

‘fisk’ is used under CC0 1.0 “flask” is used under CC0 1.0