STRspy-ing hidden variation in forensic DNA profiles using the MinION device

Courtney L. Hall[1*], Rupesh K. Kesharwani[2], Nicole R. Phillips[1], John V. Planz[1], Fritz J. Sedlazeck[2], Roxanne R. Zascavage[1]

*courtneyhall@my.unthsc.edu

[1]*Department of Microbiology, Immunology & Genetics, University of North Texas Health Science Center, Fort Worth, TX, USA*
[2]*Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA*

INTRODUCTION

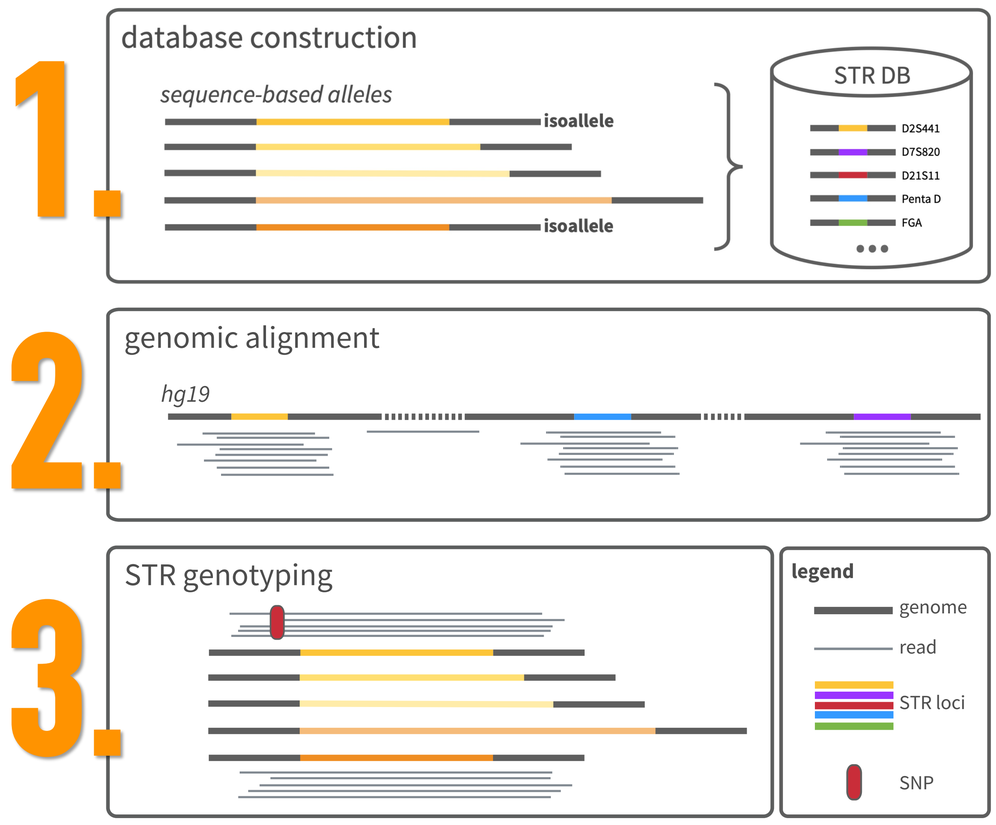
BACKGROUND. Forensic genetic analyses harness the high repeat-length variability of short tandem repeat (STR) markers for human identification. Despite the power & reliability of current typing techniques, sequence-level information within & around STRs are masked in the length-based profiles produced using capillary electrophoresis (CE). To harness the advantages of portable next generation sequencing in forensic investigations, we developed STRspy, a novel method capable of generating accurate length- & sequence-based allele designations for autosomal STRs from ONT reads.

OBJECTIVE. Here we expand upon STRspy to enable simultaneous profiling of autosomal & Y-STRs as well as flanking SNPs.

HIGHLIGHTS.

- Combined autosomal & Y-STR panels can be amplified & sequenced on the MinION device.
- STRspy achieved 100% concordance based on sequence & length across the 22 autosomal & 23 Y-STRs amplified at 30 PCR cycles.
- Isoalleles within & between samples can be resolved with ONT reads when analyzed with STRspy.
- SNPs flanking autosomal STRs were detected with >90% accuracy in the 15-cycle dataset.
- Our methods support sequencing & profiling of at least 24 amplicon libraries per MinION flow cell.

STRSPY



STRspy relies on a user-generated STR database (DB) containing sequence-based alleles for each locus of interest. We updated our STR DB to include all 22 autosomal & 23 Y-STRs amplified with the Promega PowerSeq 46GY System.

Basecalled reads are first aligned to the human reference genome. Reads overlapping loci defined in the user-provided bed file are then extracted & mapped to the STR DB.

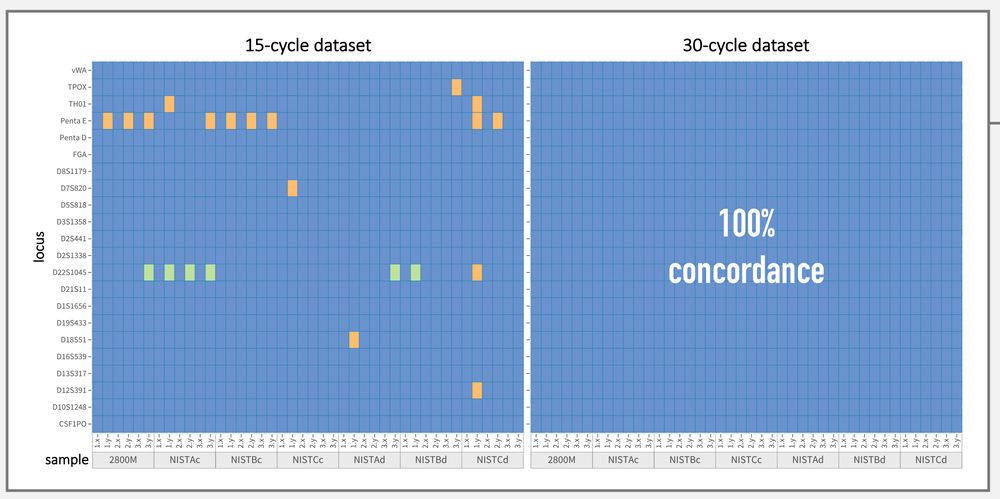
STRspy employs a user-defined threshold to report heterozygous & homozygous autosomal genotypes based on normalized read counts. The updated version of STRspy also predicts the top allele for all Y-STRs except DYS385ab. These duplicate loci have identical flanking regions & are profiled like autosomal STRs.

METHODS & RESULTS

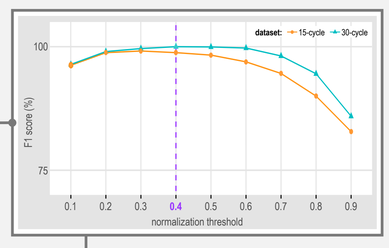
SAMPLE PROCESSING. Seven control DNAs (female n = 2; male n = 5) were amplified at 15 & 30 PCR cycles using the Promega PowerSeq 46GY System with 0.5ng of input DNA in triplicate. The multiplexing experiment was conducted using stock solutions of 30-cycle barcoded amplicons pooled to 75ng in sets of 12, 18 & 24 libraries per MinION flow cell. All sequencing data were generated on the MinION device.

DATA ANALYSIS. Basecalled reads were analyzed with the updated version of STRspy (see above). Resultant STR allele designations & flanking region SNP calls were compared to the manufacturer-validated profiles. Correct STRspy predictions were classified as true positives (blue), incorrect as false positives (green) & dropout as false negatives (orange).

autosomal STRs

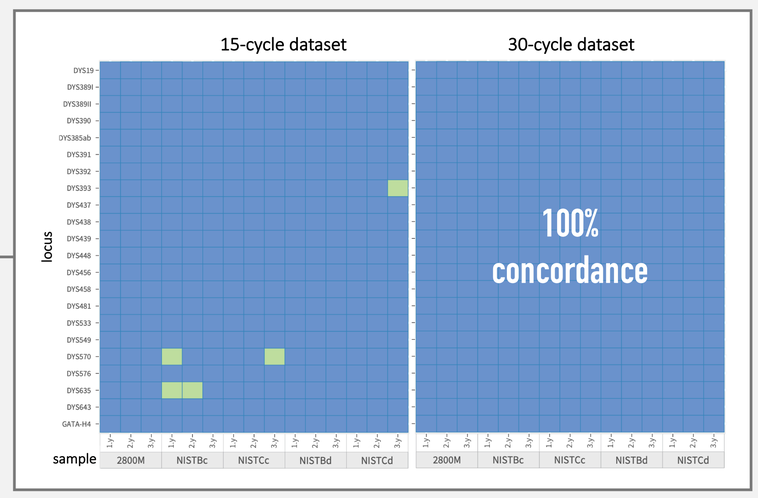


benchmarking



	autosomal 15	autosomal 30	cycles	Y 15	Y 30	
#	902	924	true positives	310	315	#
#	6	0	false positives	5	0	#
#	16	0	false negatives	0	0	#
%	98.3	100	recall	98.4	100	%
%	99.3	100	precision	98.4	100	%
%	98.8	100	F1 score	98.4	100	%

Y-STRs



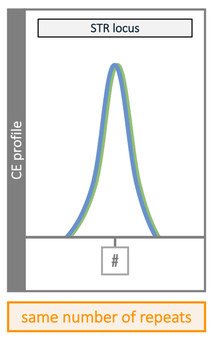
multiplexing

As expected, the number of STR aligned reads decreased with increasing multiplex size. Nevertheless, STRspy predicted the correct allele designations across all 22 autosomal & 23 Y-STRs assessed. These results demonstrate that accurate & reproducible profiles can be generated for the largest multiplex tested in forensic STR sequencing applications to date.

STRspy can...

generate CODIS-compatible allele calls.

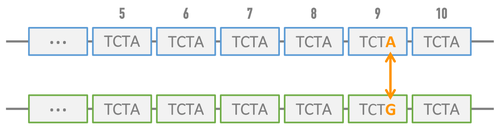
length-based designations



resolve sequence-based heterozygotes at autosomal loci.

autosomal STR

D2S441
NISTAc (10, 10)

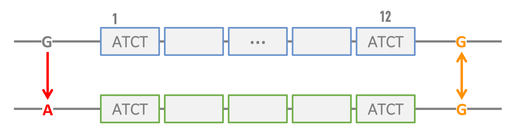


normalized read counts	30-cycle 1	30-cycle 2	30-cycle 3	15-cycle 1	15-cycle 2	15-cycle 3
[TCTA]10	0.86	1.00	0.75	1.00	1.00	0.58
[TCTA]8 TCTG TCTA	1.00	0.84	1.00	0.98	0.89	1.00

detect SNPs in autosomal STR flanking regions with high accuracy.

autosomal flanking SNPs

D5S818
NISTBd (12, 12)

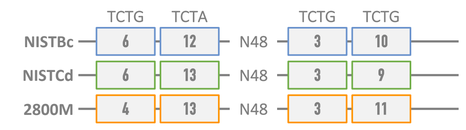


benchmarking metrics	% recall	% precision	% F1
30-cycle	92.06	74.05	82.08
15-cycle	98.41	84.05	90.66

resolve Y-STR isoalleles between samples with low coverage.

Y-STR

DYS391 II
(31)



raw read counts	30-cycle 1	30-cycle 2	30-cycle 3	15-cycle 1	15-cycle 2	15-cycle 3
NISTBc	235528	140159	224587	72	199	87
NISTCd	121459	238896	116302	76	270	168
2800M	149092	203190	149593	88	224	198

ONGOING RESEARCH. We are updating our STR database to contain all loci & alleles reported in the common autosomal & Y subdivisions of the STRSeq BioProject. This improvement will allow users to harness the most comprehensive collection of validated autosomal & Y-STRs based on NGS data with STRspy. We also aim to determine whether STRspy can be used to profile biological materials encountered in routine forensic casework (buccal swab, blood & bone). Other efforts will be geared towards expanding the current capabilities of our novel bioinformatic method for the simultaneous detection of SNPs & variation within mtDNA. In addition to traditional PCR amplification, we will assess the use of the probe-based capture methods to minimize stutter & improve profiling of severely degraded samples. These studies could enable us to achieve the most comprehensive representation of forensic genetic variation to date with the pocket-sized MinION device.

ACKNOWLEDGEMENTS. This project was funded by the National Institute of Justice (Awards 2015-DN-BX-K068 & 2018-DU-BX-0179).