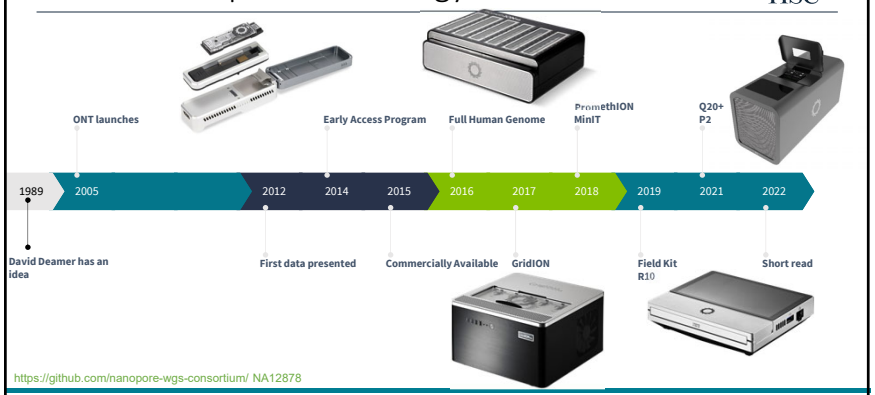
The Latest on Nanopore Sequencing in Human Identification

Roxanne R. Zascavage, Ph.D.

CL Hall, RK Kesharwani, NR Phillips, JV Planz, FJ Sedlazeck, RR Zascavage

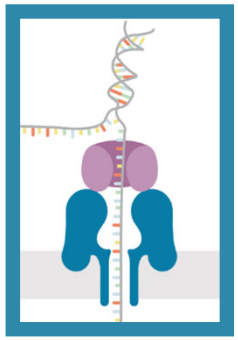
Oxford Nanopore Technology



https://github.com/nanopore-wgs-consortium/ NA12878

2

How it works



Basecalled nucleotides

CGACTCCGGTTACCAGCGCGTTGAT

Data analysis

Downstream bioinformatics pipelines

Research Aim

Evaluate the forensic applicability of the newest & smallest NGS platform

Advantages

- Portable
- Low startup cost
- Plugs into PC or laptop via USB port
- **Potential for on-site sample processing**

But…

- Relatively high error rate
- "long read" sequencer
- **Lack of forensic-specific analysis software**
- **Unable to correctly type all STR loci amplified by established multiplexes**

Where I Left You Last Year…

Sequence entire mtGenome sans enrichment

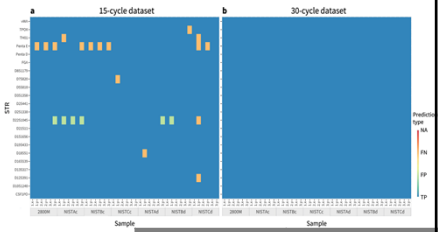
STRspy correctly profiled 22 autosomal STRs amplified at 30 PCR cycles across all samples

Concordance

- 99.6% identical

Error

- 0.4 vs. 0.3% non-amped vs amped
- All in homopolymeric stretches, except one



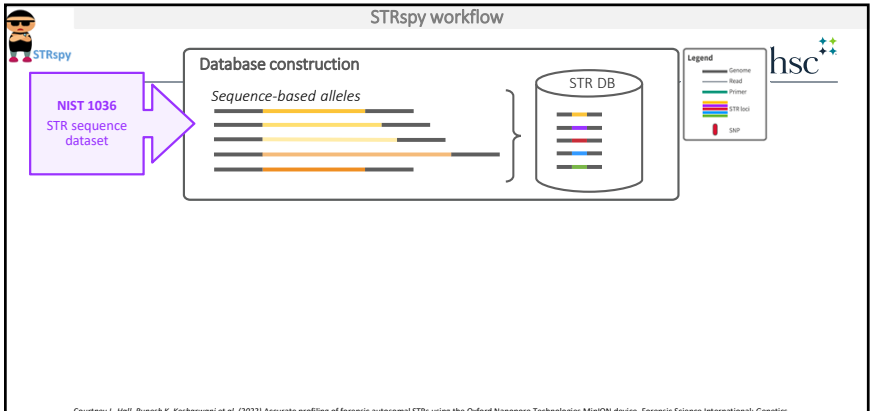
Results

SAMPLE	✓ Number of T calls	Number of C calls
HL60 non-amped	84 (75.68%)	19 (17.12%)
HL60 amped	12,798 (45.90%)	13,418 (48.12%)

sequence-based heterozygote

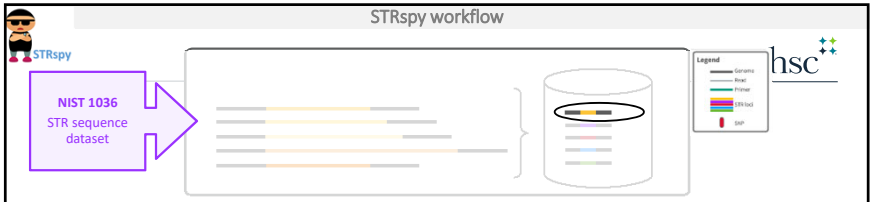
	30-cycle			15-cycle		
	1	2	3	1	2	3
[TCTA]10	0.86	1.00	0.75	1.00	1.00	0.58
[TCTA]8 TCTG TCTA	1.00	0.84	1.00	0.98	0.89	1.00

STRspy workflow

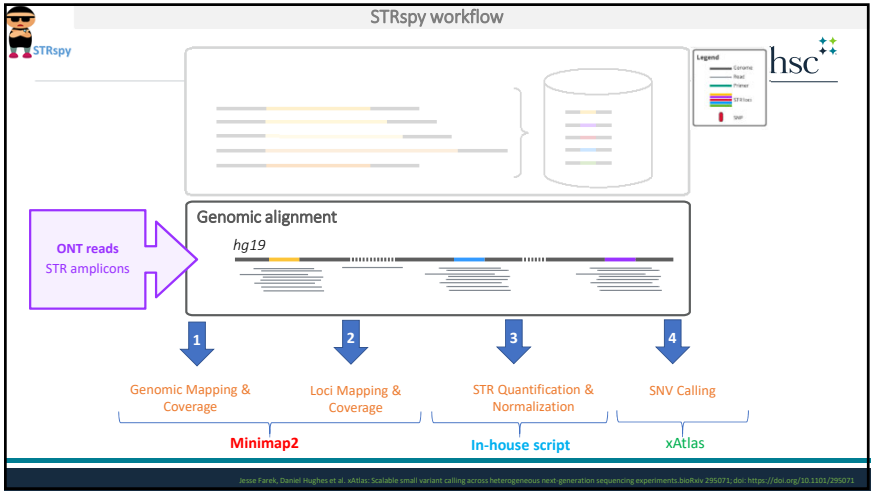


Courtney L. Hall, Rupesh K. Kesharwani et al. (2022) Accurate profiling of forensic autosomal STRs using the Oxford Nanopore Technologies MinION device, Forensic Science International: Genetics.

STRspy workflow

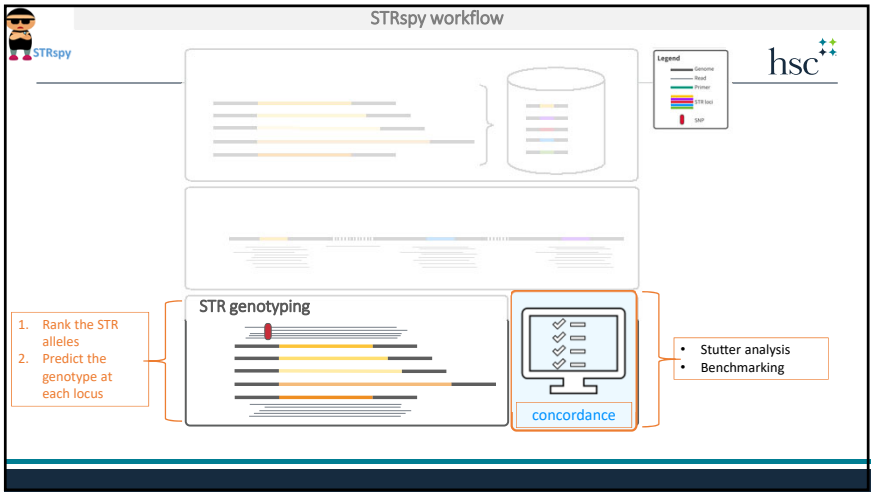


STRspy workflow



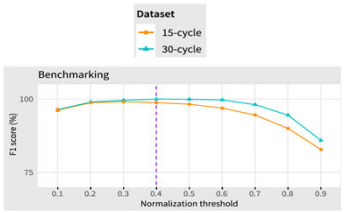
Jesse Farek, Daniel Hughes et al. xAtlas: Scalable small variant calling across heterogeneous next-generation sequencing experiments. bioRxiv 295071; doi: https://doi.org/10.1101/295071

STRspy workflow



STRspy workflow

STRspy's prediction threshold



- Shows F1 score across different normalization thresholds (0.1 to 0.9).
- The dotted lines shows the optimal threshold at both PCR cycles (15 and 30).

green = true alleles
blue = examples of n±1 stutter

30-PCR Cycle

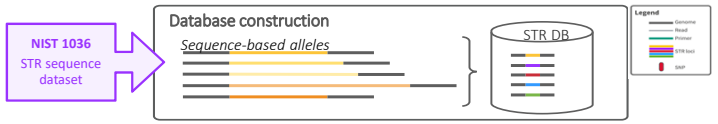
Locus	[Repeat motif]_allele designation	Raw	Norm
CSF1PO	[ATCT]10_10	203904	1
	[ATCT]11_11	29263	0.14351
	[ATCT]9_9	20664	0.10134
	[ATCT]12_12	3032	0.01487
	[ATCT]8_8	1823	0.00894
	[ATCT]9_ACCT_10	1720	0.00844
	[ATCT]7_ACCT_[ATCT]3_11	1562	0.00766
	[ATCT]15_15	560	0.00275
	[ATCT]13_13	441	0.00216
	[ATCT]8_ACCT_[ATCT]3_12	382	0.00187
	[ATCT]7_7	336	0.00165
	[ATCT]14_14	234	0.00115
	[ATCT]3_CTCT_[ATCT]3_7	66	0.00032
D10S1248	[GGAA]16_16	118646	1
	[GGAA]15_15	106719	0.89947
	[GGAA]17_17	34269	0.28883
	[GGAA]14_14	26649	0.22461
	[GGAA]18_18	9989	0.08419
	[GGAA]13_13	6328	0.05334
	[GGAA]19_19	4079	0.03438
	[GGAA]12_12	2490	0.02099
	[GGAA]9_AGAA_[GGAA]3_13	1807	0.01523
	[GGAA]11_11	658	0.00555
	[GGAA]10_10	192	0.00162
	[GGAA]8_8	88	0.00074
	[GGAA]9_9	66	0.00056

STRspy Limitations

Resolved 22 autosomal STRs

- What about Y-STRs???

Self established database



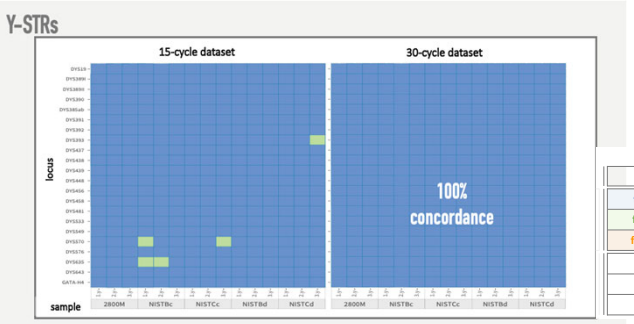
- Used Promega Powerseq 46 kit
- What about other kits and other loci?
  - Works on "best fit" model → missing loci = inaccurate calls

Nomenclature Reporting

- Inconsistencies between databases
- SNVs reported after sequencing

STRspy-ing Y-STRs

- "21" STRs vs 15 Samples (315 in total) were tested for Y-STR.



STRspy correctly profiled 22 autosomal STRs amplified at 30 PCR cycles across all samples

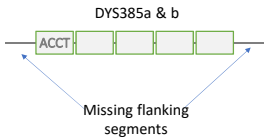
	Y	
cycles	15	30
true positives	310	315
false positives	5	0
false negatives	0	0
recall	98.4	100
precision	98.4	100
F1 score	98.4	100

https://github.com/unique379r/strspy

STRspy-ing Y-STRs

resolve Y-STR isoalleles between samples with low coverage.

- Call as biallelic loci
- Cannot distinguish definitively



Y-STR

DYS391 II (31)

NISTBc	TCTG 6	TCTA 12	N48	TCTG 3	TCTG 10
NISTCd	6	13	N48	3	9
2800M	4	13	N48	3	11

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

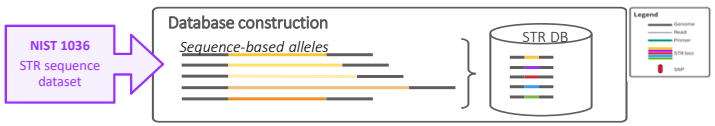
https://github.com/unique379r/strspy

STRspy Limitations

~~Resolved 22 autosomal STRs~~

- ~~What about Y-STRs???~~

Self established database



- Used Promega Powerseq 46 kit
- What about other kits and other loci/alleles?
  - Works on "best fit" model → missing alleles = inaccurate calls

Nomenclature Reporting

- SNVs reported after sequencing
- Inconsistencies between databases

STRspy Database Updates

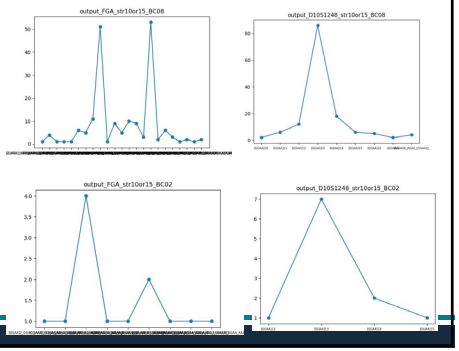
Updated the database to include all reported alleles in the STRseq BioProject

54 total loci

- 24 autosomal STRs
- 29 Y-STRs
- Sex determining amelogenin

Provide options for peak-based outputs

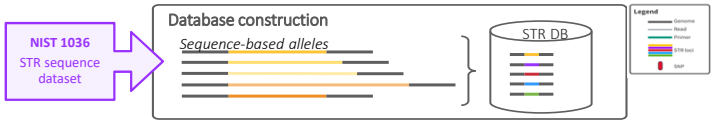
FGA D10S1248



STRspy Limitations

~~Resolved 22 autosomal STRs~~
- ~~What about Y-STRs???~~

~~Self established database~~

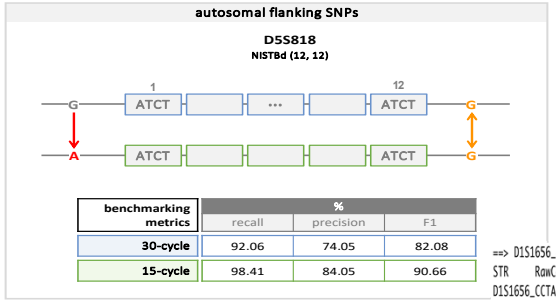


- ~~Used Promega Powerseq 46 kit~~
- ~~What about other kits and other loci/alleles?~~
  - ~~Works on "best fit" model → missing alleles = inaccurate calls~~

Nomenclature Reporting
- ~~Inconsistencies between databases~~
- SNVs reported after sequencing

Consider SNP calling adjustments…

STRspy-ing SNPs



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

--->Imputed SNP on fasta DB along with rs_id.

```
==> D1S1656_dB_30-0.5-1.fastq.minimap.sorted.bam_Allele_freqs.txt <==
STR     RawCounts       NormalizedCounts
D1S1656_CCTA_[TCTA]10_TCA_[TCTA]4_15.3_rs4847015         19099   1
D1S1656_[TCTA]13_13     12598   0.659616
```

STRspy-ing SNPs

```
>D1S1656_CCTA_[TCTA]13_14
>D1S1656_CCTA_[TCTA]13_14_rs1019813099
--
>D1S1656_CCTA_[TCTA]10_TCA_[TCTA]4_15.3
>D1S1656_CCTA_[TCTA]10_TCA_[TCTA]4_15.3_rs4847015
>D1S1656_CCTA_[TCTA]11_TCA_[TCTA]3_15.3
>D1S1656_CCTA_[TCTA]11_TCA_[TCTA]3_15.3_rs4847015
--
>D1S1656_CCTA_[TCTA]11_TCA_[TCTA]4_16.3
>D1S1656_CCTA_[TCTA]11_TCA_[TCTA]4_16.3_rs4847015
>D1S1656_CCTA_[TCTA]12_TCA_[TCTA]3_16.3
>D1S1656_CCTA_[TCTA]12_TCA_[TCTA]3_16.3_rs4847015
>D1S1656_CCTA_[TCTA]16_17
>D1S1656_CCTA_[TCTA]16_17_rs541123499
--
>D1S1656_CCTA_[TCTA]12_TCA_[TCTA]4_17.3
>D1S1656_CCTA_[TCTA]12_TCA_[TCTA]4_17.3_rs4847015
--
>D1S1656_CCTA_[TCTA]12_TCA_TCTG_[TCTA]3_17.3
>D1S1656_CCTA_[TCTA]12_TCA_TCTG_[TCTA]3_17.3_rs4847015
--
>D1S1656_CCTA_[TCTA]13_TCA_[TCTA]4_18.3
>D1S1656_CCTA_[TCTA]13_TCA_[TCTA]4_18.3_rs4847015
>D1S1656_CCTA_[TCTA]14_TCA_[TCTA]4_19.3
>D1S1656_CCTA_[TCTA]14_TCA_[TCTA]4_19.3_rs4847015
```

TMI

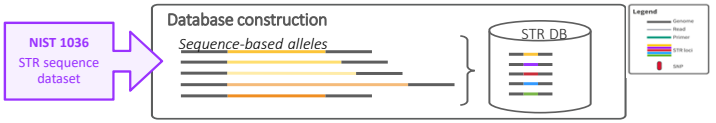
- Too many options for alignment
- Insufficient coverage to reach call thresholds

- Continue to use xAtlas post-allele calling

STRspy Limitations

~~Resolved 22 autosomal STRs~~
- ~~What about Y-STRs???~~

~~Self established database~~



- ~~Used Promega Powerseq 46 kit~~
- ~~What about other kits and other loci/alleles?~~
  - ~~Works on "best fit" model → missing alleles = inaccurate calls~~

Nomenclature Reporting
- ~~Inconsistencies between databases~~
- ~~SNVs reported after sequencing~~

ONT MinION + STRspy ready for prime time!

STRspy Ready for Prime Time?

Testing STRspy on Mock Casework Samples

22 Autosomal STRs

Sample	TP	FP	FN	Recall	Precision	F1score
Swab	726	63	11	98.507	92.015	95.150
Blood	675	147	18	97.403	82.117	89.109
Bone	201	7	0	100.00	96.635	98.289

23 Y-STRs

Sample	TP	FP	FN	Recall	Precision	F1score
Swab	185	13	0	100.00	93.434	96.606
Blood	144	42	0	100.00	77.419	87.273
Bone	-	-	-	-	-	-

STRspy is NOT Ready for Prime Time...

STRspy is NOT Ready for Prime Time...

Troubleshooting

	CE Truth Set		ONT Calls	
	1	2	1	2
D7S820	9	10	9.1	10
D13S317	11	11	12	10
TPOX	8	12	8	13

D7S820

- STRbase V2.0: 9.0
  - Does not exist in NCBI*

```
##HumanSTR-START##
Sequence attribution  :: Applied Genetics Group, NIST
STR locus name        :: D7S820
Length-based allele   :: 9.1
Minimum range bracket :: A[10]CTAT[1]CAATCTGT[1]CTAT[9]_-1T>-
Bracketed record seq. :: A[10]CTAT[1]CAATCTGT[1]CTAT[9]_-1T>-
Sequencing technology :: MiSeqFGx
Sequencing assay code :: PS
Coverage              :: >30X
Length-based tech.    :: PowerPlex Fusion 6C, 3500xl
Assembly              :: GRCh38 (GCF_000001405)
Chromosome            :: 7
Ref. seq. accession   :: NC_000007.14
Chrom. location       :: 84160149..84160346
ISFG minimum range    :: 84160200..84160281
Frequency reference   :: STRidER.online
STR locus alt name    :: D7
Historical bracketing :: [TATC]9
```

GenBank: OK330006.1

STRspy is NOT Ready for Prime Time...

Troubleshooting

3' flanking region SNP:
TATCAATCATCTATCTATCTTT...

	CE Truth Set		ONT Calls	
	1	2	1	2
D13S317	11	11	12	10

```
##HumanSTR-START##
Sequence attribution  :: Applied Genetics Group, NIST
STR locus name        :: D13S317
Length-based allele   :: 10
Minimum range bracket :: TATC[12]AATC[1]ATCT[2]
Bracketed record seq. :: TATC[12]AATC[1]ATCT[2]
Sequencing technology :: MiSeqFGx
Sequencing assay code :: FS,PS
Coverage              :: >30X
Length-based tech.    :: PowerPlex Fusion 6C, 3500xl
Assembly              :: GRCh38 (GCF_000001405)
Chromosome            :: 13
Ref. seq. accession   :: NC_000013.11
Chrom. location       :: 82147955..82148144
ISFG minimum range    :: 82148021..82148104
Frequency reference   :: STRidER.online
STR locus alt name    :: D13
Historical bracketing :: [TATC]11
```

GenBank: OK330010.1

```
##HumanSTR-START##
STR locus name        :: D13S317
Length-based allele   :: 12
Bracketed repeat      :: [TATC]13
Sequencing technology :: MiSeq FGx
Sequencing assay code :: FS
Coverage              :: >30X
Length-based tech.    :: Not reported. Allele length inferred from
                         sequence.
Assembly              :: GRCh38 (GCF_000001405)
Chromosome            :: 13
RefSeq Accession      :: NC_000013.11
Chrom. Location       :: 82147986..82148107
Repeat Location       :: 82148025..82148068
Cytogenetic Location  :: 13q31.1
 Ref. seq. accession    :: NC_000013.11
 Chrom. location        :: 82147955..82148144
 ISFG minimum range     :: 82148021..82148104
 Frequency reference    :: STRidER.online
 STR locus alt name     :: D13
 Historical bracketing :: [TATC]11
```

What about the P2?

MinION vs. P2

Research Aim: STRs, mtDNA, and SNPs in a single assay

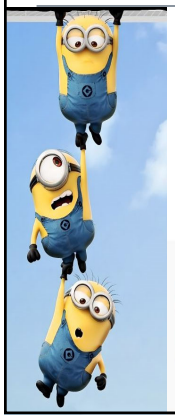
MinION

- Portable
- Up to 50Gb of data
- $1000 starter package
  - 1 flow cell and sequencing kit
- IT requirements:
  - 16GB RAM
  - CPU with 4 cores/8 threads OR
  - GPU RTX 2060 super or better
    - 8GB + GPU memory
  - 1 TB internal SSD

P2 Solo

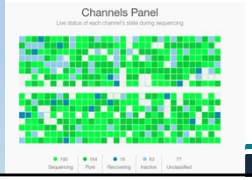
- Up to 580Gb of data
  - 100-200 Gb native gDNA reads
- $10,455 starter package
  - 8 flow cells and 2 sequencing kits
- IT requirements:
  - 16 GB DDR4+ RAM (64GB recommended)
  - GPU NVIDIA RTX minimum (NVIDIA A6000)
    - 12GB + GPU memory
  - CPU with 8 cores/16 threads (12-core/24 thread)
  - 2 TB internal SSD + 6 TB external SSD (8TB+ internal)
  - Computer cost: $7811.81
- Total cost: $18,266.81

MinION vs. P2



MinION

- Input: 75ng at 420 bases/second
  - 12.5ng X 6 samples
  - 2ng X 24 samples
- Output: 16.99 million reads



MinION vs. P2

P2

- Input: 60ng – 150ng
  - 10ng X 6 samples at 260 bases/second
  - 25ng X 2 samples + 50ng X 2 samples at 420 bases/second

Sample	Input (ng)	Total Reads	Genomic Mapping	Unmapped	STR Mapping
Blank	0	934942	4710	930232	0
nistA_1	10	1237652	1017418	220234	16
nistA_2	10	2755005	2126670	628335	29
nistB_1	10	1299002	661197	637805	6
nistB_2	10	3171316	2708639	462677	30
nistC_1	10	1642393	1222958	419435	27
nistC_2	10	1884020	1456143	427877	34
TOTAL	60	**12,924,330**	9,197,735	3,726,595	142

Sample	Input (ng)	Total Reads	Genomic Mapping	Unmapped	STR Mapping
2800m_1	25	760841	375691	385150	8
2800m_2	25	1133423	911414	222009	8
2800m_1	50	996406	775651	220755	10
2800m_2	50	2076524	1616491	460033	19
TOTAL	150	**4,967,194**	3,679,247	1,287,947	45

STR Mapping: 2800M

- 20/50 STR loci
- TPOX = 6 reads total
  - 3 from 1 25ng input run
  - 3 from 1 50ng input run
- CSF1PO = 5 reads total

MinION vs. P2

Troubleshooting and Conclusions

Higher DNA input is not necessarily better for yield

- Seen this before
- Short fragments clog pores
  - 200 ng input on MinION = 675.77k reads vs. 16.99 million from 75ng

P2 doesn't provide improvements over minION

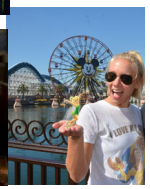
- Lower yield without more mapped reads
- Could be due to lack of protocol development

What's Next?

- P2 Improvements
- STR, SNP, and Mito Multiplex Assay
- Challenged Samples
- Portability
- STRspy development: GUI Interface
- STRspy development: Database updates

STRspy

Acknowledgements



School of Biomedical Sciences

Team

- Courtney Hall
- Nicole Phillips
- Roxanne Zascavage
- Bupe Kapema

Baylor College of Medicine HGSC Human Genome Sequencing Center

- Fritz Sedlazeck
- Rupesh Kesharwani

Funding

NIJ National Institute of Justice

STRENGTHEN SCIENCE. ADVANCE JUSTICE.

Questions?

	No barcodes	24 barcodes	48 barcodes	96 barcodes
Flow cell price	$500	$500	$500	$500
Library price	$99	$99	$99	$99
Barcodes price	-	$75	$150	$300
Price per sample	$599	$28	$15	$9.36