Characterization of NIST Standard Reference Materials by Next Generation Sequencing

National Institute of Standards and Technology
Technology Administration, U.S. Department of Commerce

Email:
Kevin.Kiesler@nist.gov

Kevin M. Kiesler and Peter M. Vallone

U.S. National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, MD 20899-8314, USA

P-299

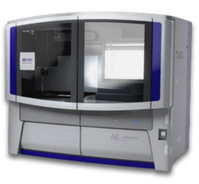
The National Institute of Standards and Technology (NIST) offers certified Standard Reference Materials (SRMs) for laboratories performing DNA-based human identity testing. Developed by the Applied Genetics group at NIST, forensic DNA SRMs are well characterized for relevant markers such as autosomal and Y-chromosomal short tandem repeats (STRs) and mitochondrial DNA (mtDNA) sequence. NIST's forensic SRMs are required by the FBI Quality Assurance Standards for calibration of DNA typing instruments and procedures. Characterization of the SRMs thus far has been performed by capillary gel electrophoresis fragment-based genotyping analysis and/or Sanger sequencing. Ultra high-throughput sequencing, or next-generation sequencing (NGS), offers an opportunity to use an orthogonal method to further characterize DNA-based reference materials and thereby increase the confidence in the certified values of the materials. The additional sensitivity of NGS methods may increase the information content of the SRMs through the characterization of new features (e.g. SNPs, insertion-deletions, minor variants) within the markers' sequences which are beyond the ability of fragment-based analysis or Sanger sequencing to detect. This work will ultimately lead to certified reference materials for laboratories wishing to implement and validate NGS technology in the field of forensics. Characterization of NIST forensic SRMs 2392 and 2392-I for mtDNA sequencing has been performed on several NGS platforms. Here we present results of NGS mtDNA sequencing concordance with certified values determined by the Sanger method and preliminary data on sequencing of SRM 2391c for STR typing.

Next Generation Sequencing

Several technologies for high-throughput DNA sequencing that provide high coverage have become available in the past decade. These platforms use different approaches to generate several thousand to billions of sequences per run. The highest output instruments can generate sequence at very low cost per base, but run times may be measured in weeks. These instruments are best suited to sequencing large genomes.

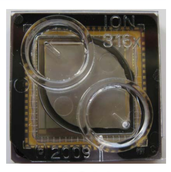
More recently small bench-top systems with the scale, cost, and run times amenable to the forensic laboratory have been introduced by Life Technologies (PGM), Illumina (MiSeq), and Roche/454 (454 jr.). These bench-top systems are priced at approximately $80 K to $125 K and cost per run is between $200 and $1100. Sequencing output ranges from 10 Mb to 1500 Mb per run and run times can take between 3 hours and 27 hours*. Multiple samples can be analyzed in a single run through the use of DNA barcodes introduced during the library construction phase. The number of samples analyzed will depend on the instrument's sequencing output and the user's desired level of coverage. *Output, cost, and run time estimates are from manufacturers' technical specifications

The PGM and 5500 instruments are currently available at NIST



5500 SOLiD

Ion Torrent PGM



PGM
Ion 316 chip

The Illumina MiSeq will be installed this fall at NIST

MiSeq

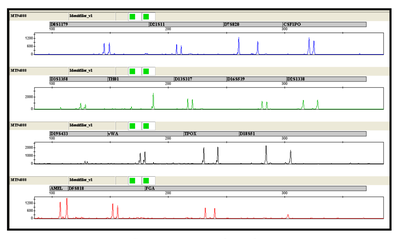
We have also outsourced mitochondrial sequencing to EdgeBio on the PGM and MiSeq platforms.
http://www.edgebio.com/

NIST Standard Reference Materials (SRM)

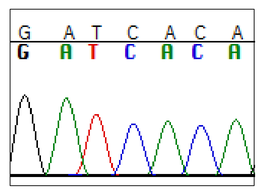
NIST offers for sale Standard Reference Materials (SRMs) for mitochondrial DNA sequencing and STR typing. SRM 2391c is a Certified Reference Material used in quality assurance procedures for validating STR loci measurements. SRM 2392 and 2392-I are Certified Reference Materials used in quality assurance procedures for validating mitochondrial DNA sequencing. These materials are primarily characterized through a combination of Sanger sequencing (2392, 2392-I) fragment length analysis (2391c).

SRMs 2391c, 2392, and 2392-I have been further characterized, through the use of NGS technology, providing additional information (full sequencing of STR motifs and surrounding variants) and a higher degree of sensitivity (increased coverage of the mitochondrial genome). This is being performed to support the emerging use of high-throughput sequencing for forensic applications.

By characterization of the SRMs using multiple NGS platforms we intend to limit any potential bias of using one specific NGS platform or a single computational assembly pipeline.



SRM 2391c → STR analysis



SRM 2392 and 2392-I → mtDNA sequencing

Goals: To increase confidence in the existing certified values of NIST SRMs and to expand the quantity of information available for both mitochondrial DNA sequence and STR repeat internal sequence variation through the use of next generation sequencing technology.

Methods for Next Generation Sequencing of Mitochondrial DNA

- The entire mtDNA genome was selectively amplified using a set of 12 PCR primers (see Levin *et al*. 2003) designed to produce amplicons ranging in size from 850 to 1600 bp which overlap slightly, to allow for full mtDNA sequence coverage.
- PCR amplicons were purified using the ExcelaPure UF 96-Well Purification Kit and quantified on a NanoDrop 2000 spectrophotometer. Amplicons were pooled on an equal mass ratio and libraries were generated and sequenced on the **PGM** and **5500** platforms at NIST. Additional runs on **PGM** and **MiSeq** platforms were outsourced to EdgeBio (Gaithersburg, MD, USA) and to Beckman Coulter Genomics (Danvers, MA, USA) on the **HiSeq 2000**.
- Resulting sequence was pre-processed using manufacturer's platform-specific software and sequence variants were identified using NextGENe version 2.3.3 software (Softgenetics, State College, PA, USA).
- Variants are reported as polymorphisms relative to the revised Cambridge Reference Sequence (rCRS) (Anderson *et al.* 1999). **Novel heteroplasmic sites,** previously undetected by Sanger sequencing, were identified in SRM 2392 Component B at positions **1393** and **7861** (**highlighted green** in Table 1). Average sequencing coverage ranged from **≈ 200x to 50,000x** depending on the platform and level of multiplexing.

Methods for Next Generation Sequencing of STR Loci

- STR loci were amplified in singleplex PCR reactions using PCR primers designed to bind outside known commercial primer binding sites and conditions described in Kline *et al*. 2011.
- PCR amplicons (**ranging from ≈250 to 500 bp**) were purified using the QuickStep 2 96-Well Purification Kit (Edge Biosystems) and quantified on a Bioanalyzer DNA12000 chip.
- Amplicons were pooled on an equimolar ratio and library construction was performed using the IonXpress Plus Fragment Library Kit. Libraries were sequenced on the PGM platform using both the Ion One Touch 200 v2 Template Kit and Ion PGM Sequencing 200 Kit v2 or the Ion PGM Template OT2 400 Kit and Ion PGM Sequencing 400 Kit in two separate sequencing runs.
- Resulting sequence data was pre-processed using Torrent Suite version 3.62 software then analyzed using NextGENe version 2.3.3 software (Softgenetics) using a custom reference file containing information based on repeat structures and STR loci accession numbers from ***Appendix 1 of Advanced methods in forensic DNA typing: methodology*** (Butler 2012). Allele calls were performed by examining the alignment maps and assigning the alleles for which the highest number of sequencing reads had mapped. Mapped alleles were inspected for variants in the repeat motifs and flanking sequence.

Table 1: Variants reported as differences relative to the rCRS for (a) SRM 2392 Component A (CHR) (b) SRM 2392 Component B (9947A) and (c) SRM 2392-I and (d) novel heteroplasmic sites in SRM 2392 Component B identified by next generation sequencing.

a

Nucleotide Position	rCRS Reference Sequence	SRM 2392 Component A Sanger Call	NGS Consensus
64	C	C/T	C/T
73	A	G	G
195	T	C	C
204	T	C	C
207	G	A	A
263	A	G	G
309.1		C	
315.1		C	
709	G	A	A
750	A	G	G
1438	A	G	G
1719	G	A	A
2706	A	G	G
4769	A	G	G
4929	C	T	T
5186	A	G	G
6221	T	C	C
6371	C	T	T
6791	A	G	G
7028	C	T	T
8503	T	C	C
8860	A	G	G
11719	G	A	A
11878	T	C	C
12612	A	G	G
12705	C	T	T
13708	G	A	A
13966	A	G	G
14470	T	C	C
14766	C	T	T
15326	A	G	G
16183	A	C	C
16189	T	C	C
16193.1		C	
16223	C	T	T
16278	C	T	T
16519	T	C	C

b

Nucleotide Position	rCRS Reference Sequence	SRM 2392 Component B Sanger Call	NGS Consensus
93	A	G	G
195	T	C	C
214	A	G	G
263	A	G	G
309.1		C	
309.2		C	
315.1		C	
750	A	G	G
1393	G	G	G/A
1438	A	G	G
4135	T	C	C
4769	A	G	G
7645	T	C	C
7861	T	C	T/C
8448	T	C	C
8860	A	G	G
9315	T	C	C
13572	T	C	C
13759	G	A	A
15326	A	G	G
16311	T	C	C
16519	T	C	C

c

Nucleotide Position	rCRS Reference Sequence	SRM 2392-I Sanger Call	NGS Consensus
73	A	G	G
150	C	T	T
152	T	C	C
263	A	G	G
295	C	T	T
315.1		C	
489	T	C	C
750	A	G	G
1438	A	G	G
2706	A	G	G
4216	T	C	C
4769	A	G	G
5228	C	G	G
5633	C	T	T
7028	C	T	T
7476	C	T	T
8860	A	G	G
10172	G	A	A
10398	A	G	G
11251	A	G	G
11719	G	A	A
12071	T	C/T	C/T
12612	A	G	G
13708	G	A	A
14569	G	A	A
14766	C	T	T
15257	G	A	A
15326	A	G	G
15452	C	A	A
15812	G	A	A
16069	C	T	T
16193	C	T	T
16278	C	T	T
16362	T	C	C

- In this work 'NGS consensus' sequence consists of the combining data from three PGM runs, 5500, MiSeq and HiSeq 2000.
- The sequence that agrees across all platforms is considered to be of high confidence and reported in the tables.
- HV1 and HV2 C-stretch regions were not properly called using commercial software packages (highlighted in yellow).

d

Position 1393	
Reference G	Variant A
82 % (± 2 %)	18 % (± 2 %)
Position 7861	
Reference T	Variant C
14 % (± 3 %)	86 % (± 3 %)

Heteroplasmies not reliably detected using Sanger sequencing

Conclusion:

- The consensus data from the four NGS platforms for the mitochondrial SRMs agree with Sanger sequencing data.
- **G/A heteroplasmy at 1,393 confirmed**
- **T/C heteroplasmy at 7,861 confirmed**
- C insertions and deletions are issues (assemblers/variant callers)
- The majority of false positives are of low abundance and not reproducible across platforms.

Information in SRM 2392 certificate will be updated

Table 2: STR repeat structures identified by next generation sequencing of (a) SRM 2391c Component A (b) SRM 2391c Component B and (c) SRM 2391c Component C

a

2391c Component A				
Locus	Certified Value	NGS Value	Repeat Structure - Allele 1	Repeat Structure - Allele 2
D2S1338	18,23	TBD	To Be Determined	To Be Determined
D3S1358	15,16	15, 16	TCTA[TCTG]$_2$[TCTA]$_{12}$	TCTA[TCTG]$_3$[TCTA]$_{12}$
D5S818	11,12	11,12	[AGAT]$_{11}$	[AGAT]$_{12}$
D7S820	11,11	11,11	[GATA]$_{11}$	[GATA]$_{11}$
D8S1179	13,14	13,14	[TCTA]$_{13}$	[TCTA]$_2$TCTG[TCTA]$_{11}$
D13S317	8,8	8,8	[TATC]$_8$	[TATC]$_8$
D16S539	10,11	10,11	[GATA]$_{10}$	[GATA]$_{11}$
D18S51	12,15	12,15	[AGAA]$_{12}$	[AGAA]$_{15}$
D19S1443	13,14	TBD	To Be Determined	To Be Determined
D21S11	28,32.2	28,32.2	[TCTA]$_4$[TCTG]$_6$ {[TCTA]$_3$TA[TCTA]$_3$TCA[TCTA]$_2$TCCATA}[TCTA]$_{10}$	[TCTA]$_5$[TCTG]$_6$ {[TCTA]$_3$TA[TCTA]$_3$TCA[TCTA]$_2$TCCATA} [TCTA]$_{12}$ TA TCTA
CSF1PO	10,10	10,10	[AGAT]$_{10}$	[AGAT]$_{10}$
FGA	21,23	21,23	[TTTC]$_3$TTTT TTCT[CTTT]$_{13}$CTCC[TTCC]$_2$	[TTTC]$_3$TTTT TTCT[CTTT]$_{15}$CTCC[TTCC]$_2$
Penta D	9,13	9,13	[AAAGA]$_9$	[AAAGA]$_{13}$
Penta E	5,10	5,10	[AAAGA]$_5$	[AAAGA]$_{10}$
TH01	8,9.3	8,9.3	[AATG]$_8$	[AATG]$_6$ATG[AATG]$_3$
TPOX	8,8	8,8	[AATG]$_8$	[AATG]$_8$
VWA	18,19	18,19	TCTA[TCTG]$_4$[TCTA]$_{13}$	TCTA[TCTG]$_4$[TCTA]$_{14}$
AMEL	X, X	X, X	No Polymorphisms Observed	No Polymorphisms Observed

b

2391c Component B				
Locus	Certified Value	NGS Value	Repeat Structure - Allele 1	Repeat Structure - Allele 2
D2S1338	17,17	TBD	To Be Determined	To Be Determined
D3S1358	15,19	15, 19	TCTA[TCTG]$_3$[TCTA]$_{11}$	TCTA[TCTG]$_3$[TCTA]$_{15}$
D5S818	12,13	12,13	[AGAT]$_{12}$	[AGAT]$_{13}$
D7S820	10,10	10,10	[GATA]$_{10}$	[GATA]$_{10}$
D8S1179	10,13	10,13	[TCTA]$_{10}$	[TCTA]$_{13}$
D13S317	9,12	9,12	[TATC]$_9$	[TATC]$_{12}$
D16S539	10,13	10,13	[GATA]$_{10}$	[GATA]$_{13}$
D18S51	13,16	13,TBD	[AGAA]$_{13}$	To Be Determined
D19S1443	16,16.2	TBD	To Be Determined	To Be Determined
D21S11	32,32.2	32, 32.2	**[TCTA]$_4$[TCTG]$_6$ {[TCTA]$_3$TA[TCTA]$_3$TCA[TCTA]$_2$TCCATA}[TCTA]$_{14}$**	[TCTA]$_5$[TCTG]$_6$ {[TCTA]$_3$TA[TCTA]$_3$TCA[TCTA]$_2$TCCATA}[TCTA]$_{12}$ TA TCTA
CSF1PO	10,11	10,11	[AGAT]$_{10}$	[AGAT]$_{11}$
FGA	20,23	20,23	[TTTC]$_3$TTTT TTCT[CTTT]$_{12}$CTCC[TTCC]$_2$	[TTTC]$_3$TTTT TTCT[CTTT]$_{15}$CTCC[TTCC]$_2$
Penta D	8,12	8,12	[AAAGA]$_8$	[AAAGA]$_{12}$
Penta E	7,15	7,15	[AAAGA]$_7$	[AAAGA]$_{15}$
TH01	6,9.3	6,9.3	[AATG]$_6$	[AATG]$_6$ATG[AATG]$_3$
TPOX	8,11	8,11	[AATG]$_8$	[AATG]$_{11}$
VWA	17,18	17,18	TCTA[TCTG]$_4$[TCTA]$_{12}$	TCTA[TCTG]$_4$[TCTA]$_{13}$
AMEL	X, Y	X, Y	No Polymorphisms Observed	No Polymorphisms Observed

c

2391c Component C				
Locus	Certified Value	NGS Value	Repeat Structure - Allele 1	Repeat Structure - Allele 2
D2S1338	19,19	TBD	To Be Determined	To Be Determined
D3S1358	16,18	16,18	TCTA[TCTG]$_3$[TCTA]$_{12}$	TCTA[TCTG]$_3$[TCTA]$_{14}$
D5S818	10,11	10,11	[AGAT]$_{10}$	[AGAT]$_{11}$
D7S820	10,12	10,12	[GATA]$_{10}$	[GATA]$_{12}$
D8S1179	10,17	10,17	[TCTA]$_{10}$	**[TCTA]$_2$[TCTG]$_1$[TCTA]$_{14}$**
D13S317	11,11	12,12	**[TATC]$_{12}$ Del ATCA 6 bp ds**	**[TATC]$_{12}$ Del ATCA 6 bp ds**
D16S539	10,10	10,10	[GATA]$_{10}$	[GATA]$_{10}$
D18S51	16,19	TBD,19	To Be Determined	[AGAA]$_{19}$
D19S1443	13.2,15.2	TBD	To Be Determined	To Be Determined
D21S11	29,30	29,30	[TCTA]$_4$[TCTG]$_6$ {[TCTA]$_3$TA[TCTA]$_3$TCA[TCTA]$_2$TCCATA}[TCTA]$_{11}$	[TCTA]$_6$[TCTG]$_5$ {[TCTA]$_3$TA[TCTA]$_3$TCA[TCTA]$_2$TCCATA}[TCTA]$_{11}$
CSF1PO	10,12	10,12	[AGAT]$_{10}$	[AGAT]$_{12}$
FGA	24,26	24,26	[TTTC]$_3$TTTT TTCT[CTTT]$_{16}$CTCC[TTCC]$_2$	[TTTC]$_3$TTTT TTCT[CTTT]$_{18}$CTCC[TTCC]$_2$
Penta D	10,11	10,11	[AAAGA]$_{10}$	[AAAGA]$_{11}$
Penta E	12,13	12,13	[AAAGA]$_{12}$	[AAAGA]$_{13}$
TH01	6,8	6,8	[AATG]$_6$	[AATG]$_8$
TPOX	11,11	11,11	[AATG]$_{11}$	[AATG]$_{11}$
VWA	16,18	16,18	TCTA[TCTG]$_4$[TCTA]$_{11}$	TCTA[TCTG]$_4$[TCTA]$_{13}$
AMEL	X, Y	X, Y	No Polymorphisms Observed	No Polymorphisms Observed

Conclusion:

- Characterization of STR loci with next generation sequencing offers the potential for discerning variants of STR repeat motifs not accessible by fragments analysis.
- The bolded sequence in Table 2 reflect novel motifs not found in ***Appendix 1 of Advanced methods in forensic DNA typing: methodology*** (Butler 2012). These motifs will be added to our STR allele reference files used in sequence assembly**.**
- **The data on this poster is not certified until an updated Certificate of Analysis for SRM 2391c is released.**

Table 3: Polymorphisms in flanking region sequence identified by next generation sequencing

Locus	SRM 2391c Component	Allele	Flanking Region Variant (us = upstream, ds = downstream)	Confirmed by Sanger
D5S818	A	12	T→C 13 bp us of the repeat	Yes
D5S818	B	13	T→C 13 bp us of the repeat	Yes
D5S818	B	13	G→T 4 bp ds of the repeat	Yes
D5S818	C	10	T→C 13 bp us of the repeat	Yes
D5S818	C	11	T→C 13 bp us of the repeat	Yes
D7S820	C	10	T→G 65 bp ds of the repeat	Yes
D13S317	C	11	A→C 115 bp ds of the repeat	Yes
D16S539	A	10	A→C 16 bp ds of the repeat	Yes
D16S539	A	10	C→A 95 bp us of the repeat	Yes
D16S539	A	11	C→A 95 bp us of the repeat	Yes
D16S539	B	10	C→A 95 bp us of the repeat	Yes
D16S539	C	10	C→A 95 bp us of the repeat	Yes
Penta E	A	10	G→A 123 bp us of the repeat	Yes
Penta E	A	10	A→G 268 bp us of the repeat	Yes
Penta E	A	10	A→C 280 bp us of the repeat	Yes
Penta E	B	7	G→A 123 bp us of the repeat	Yes
Penta E	B	7	A→G 268 bp us of the repeat	Yes
Penta E	B	7	A→C 280 bp us of the repeat	Yes
Penta E	B	15	G→A 123 bp us of the repeat	Yes
Penta E	B	15	A→G 268 bp us of the repeat	Yes
Penta E	B	15	A→C 280 bp us of the repeat	Yes
TPOX	A	8	T→G 148 bp ds of the repeat	Yes
TPOX	B	8	T→G 148 bp ds of the repeat	Yes

Conclusion:

- The variants listed in the Table 3 lie in the sequence flanking the STR region. These have been confirmed by the use of Sanger sequencing (**see poster 294**).
- The flanking sequence variants characterized in this work are constrained by the size of the original PCR amplicon generated (see Kline et al. 2011).

Future work for STRs:

- Complete characterization of all loci included in the SRM 2391c certificate (Y STR loci, new U.S. loci candidates, and additional autosomal STRs).
- Confirm STR sequencing results on the MiSeq platform.
- Assess NGS performance with Component D (a 3:1 mixture of A:C).
- Assess the need to characterize Components E and F (cells on a paper substrate).
- Investigate more efficient and user friendly analysis packages for STR allele calling.
- Characterize the SRM 2391c for commercially available NGS SNP or STR kits.

Future work for mtDNA:

- Experiments for setting a variant calling threshold
- Evaluate a three amplicon approach for mitochondrial DNA enrichment
- Sequence the mitoSRMs on the Pacific Biosciences platform (Collaboration with Children's National Medical Center, Washington, D.C)

References:

Andrews R.M., Kubacka I., Chinnery P.F., Lightowlers R.N., Turnbull D.M., and Howell N. (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nature Genetics* 23 **(2)**, 147.

Butler, J.M. (2012) Advanced topics in forensic DNA typing: methodology. Waltham, MA: Elsevier Inc.

Kline, M.C., Hill, C.R., Decker, A.E., and Butler, J.M. (2011) STR sequence analysis for characterizing normal, variant, and null alleles. *Forensic Science International: Genetics* **5**, 329-332.

Levin, B.C., Holland, K.A., Hancock, D.A., Coble, M., Parsons, T.J., Kienker, L.J., Williams, D.W., Jones, M.P., and Richie, K.L. (2003) Comparison of the complete mtDNA genome sequences of human cell lines - HL-60 and GMIO742A - from individuals with pro-myelocytic leukemia and leber hereditary optic neuropathy, respectively, and the inclusion of HL-60 in the NIST human mitochondrial DNA standard reference material - SRM 2392-I. Mitochondrion 2: 387-400.

Financial disclosure: This work was supported by funding from the FBI Laboratory and Biometrics Center of Excellence: Forensic DNA Typing as a Biometric tool.

Poster available for download from STRBase
http://www.cstl.nist.gov/biotech/strbase/pub_pres/KieslerISFG2013poster.pdf