Characterization of Candidate Reference Materials by NGS

Workshop: Considerations for Implementing Next Generation Sequencing (NGS) Technologies Into a Forensic Laboratory
68th Annual AAFS Meeting
Las Vegas, NV
February 23, 2016

Peter M. Vallone, Ph.D.
Applied Genetics Group
Biomolecular Measurement Division

NIST
National Institute of Standards and Technology
U.S. Department of Commerce

Outline

- **Standard Reference Materials**
- Next Generation Sequencing
- Current and future characterization plans

Role of Standards

An SRM is prepared and used for three main purposes

- To help develop accurate methods of analysis
- To calibrate measurement systems used to facilitate exchange of goods, institute quality control, determine performance characteristics, or measure a property at the state-of-the-art limit
- To ensure the long-term adequacy and integrity of measurement quality assurance programs

9.5.5 The laboratory shall check its DNA procedures annually or whenever substantial changes are made to a procedure against an appropriate and available NIST standard reference material or standard traceable to a NIST standard.

http://www.nist.gov/srm/definitions.cfm

Certified Reference Material

- Is the same as Standard Reference Material
  - NIST name for CRM; SRM™
- Reference material characterized by a metrologically valid procedure for one or more specified properties, accompanied by a certificate that provides the value of the specified property, its associated uncertainty, and a statement of metrological traceability

Sequence, genotype, quantity/copies of DNA

Certified, Reference & Information Values

Certified Value
- NIST has highest confidence in accuracy
- All known/suspected sources of bias investigated/taken into account

Two or more methods e.g. Sanger sequencing AND genotyping with multiple primer sets

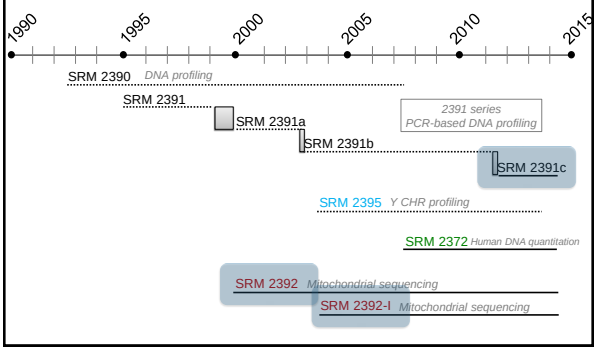
Reference Value
- Best estimate of true value
- All possible sources of bias NOT fully investigated by NIST

Genotyping with only two sets of primers

Information Value
- Of interest and use to SRM user
- Insufficient information available to assess uncertainty of value

Genotyping with only one set of primers

NIST Forensic SRM Timeline



SRM 2391c: PCR-Based DNA Profiling Standard

Primary use: STR genotyping

Certified values
25 aSTRs
29 Y-STRs

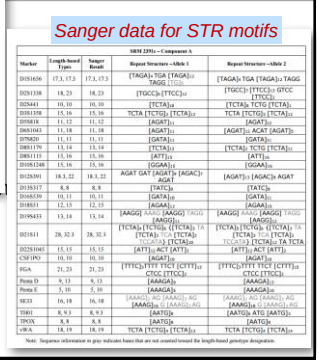
Reference values
26 aSTRs

Informational values
1 aSTR
12 X-STRs
30 InDel Loci

2391c: Profiling Standard

includes new US core loci

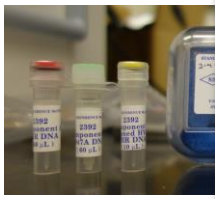
Sanger data for STR motifs



SRM 2392 and 2392-I Mitochondrial DNA Sequencing (Human)

Certified values: Sequence for the mitochondrial genome as determined by Sanger methods

SRM 2
Mitochondrial D



Certified values: S
genome as dete

Table 1. Certified Human mtDNA Sequence Differences from the Revised Cambridge Reference Sequence for SRM 2932 Component CHR.

Site	rCRS	CHR	Comment
64	C	Y = CT	C/T heteroplasmy
73	A	G	
195	T	C	
204	T	C	
207	G	A	
263	A	G	
309.1		C	insertion
315.1		C	insertion
709	G	A	
750	A	G	
1438	A	G	
1719	G	A	
2706	A	G	
3107	C		deletion
4769	A	G	
4929	C	T	
5186	A	G	
6221	T	C	
6371	C	T	
6791	A	G	
7028	C	T	
8503	T	C	
8860	A	G	
11719	G	A	
11878	T	C	
12612	A	G	
12705	C	T	
13708	C	A	
13966	A	G	
14470	T	C	
14766	C	T	
15326	A	G	
16183	A	C	
16189	T	C	
16193.1		C	insertion
16223	C	T	
16278	C	T	
16519	T	C	

NGS and Reference Materials

- A reference material can be used to confirm or validate a new technique (NGS)
  - Compare to Sanger sequence base calls or allele length/sequence
  - Sanger data for STRs in SRM 2391c
- The new technique can also be used to further characterize the reference material
  - Once the sources of bias are understood

Outline

- Standard Reference Materials
- **Next Generation Sequencing**
- Current and future characterization plans

NGS (or MPS)

- Relatively new technology
- Higher degree of
  - Coverage; markers, clonal amplification
- Easier than Sanger *?*
- Bioinformatics is an important aspect
- Commercial kits are now available for STR/SNP typing
- *Is Sanger sequencing still the gold standard?*

Current NGS Platforms in use at NIST

- Illumina
  - **MiSeq/FGx**
  - HiSeq 2000/2500
- Life Technologies
  - **Ion Torrent PGM**

Current NGS software tools in use at NIST

illumina Universal Analysis Software

- Torrent Suite (PGM)
- Lisa Borsuk (NIST) – custom tools

CLC bio A QIAGEN Company

Galaxy

https://galaxyproject.org/

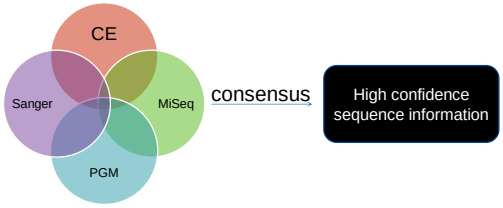
NextGENe

STRait Razor v2.0: The improved STR Allele Identification Tool – Razor

David H. Warshauer, Jonathan L. King, Bruce Budowle

Multiple characterization methods

- Use of multiple methods to obtain a consensus sequence for the SRMs
  - Identify and reduce error or bias
  - Identify and control for bias in a specific chemistry and/or informatics pipeline



Outline

- Standard Reference Materials
- Next Generation Sequencing
- **Current and future characterization plans**

SRM 2392 & 2392-I

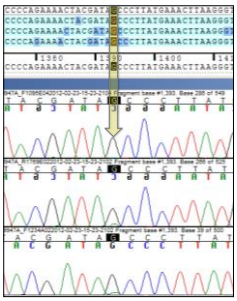
Mitochondrial DNA Sequence

- Obvious candidate for NGS characterization
- Compare and confirm sequence calls made with Sanger methods
- High coverage may provide further characterization of heteroplasmy

SRM 2392 & 2392-I

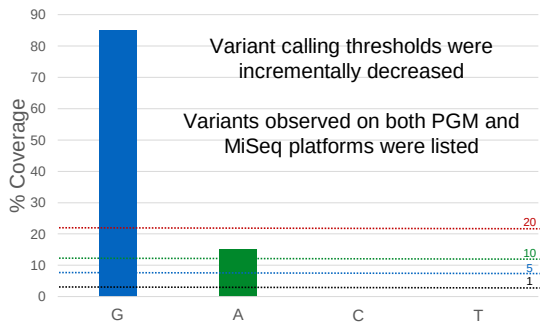
Mitochondrial DNA Sequence

- PGM & MiSeq analysis
  - Variants from rCRS were confirmed
  - Concordance across platforms
  - Lower level heteroplasmies were observed
    - Ambiguous by Sanger sequencing
  - 20, 10, 5, 1% SNP calling thresholds examined



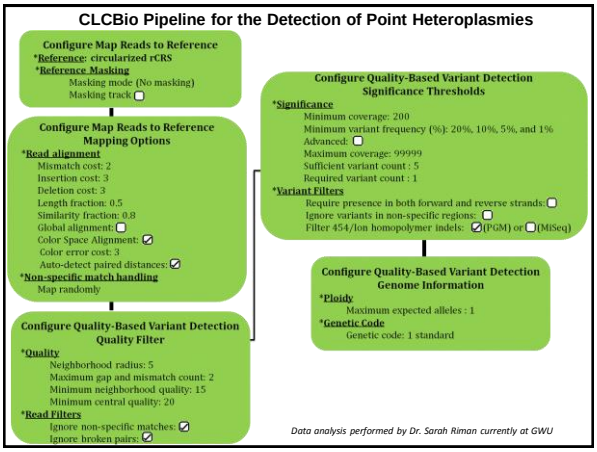
Site 1393 (G/A)

Varying Thresholds

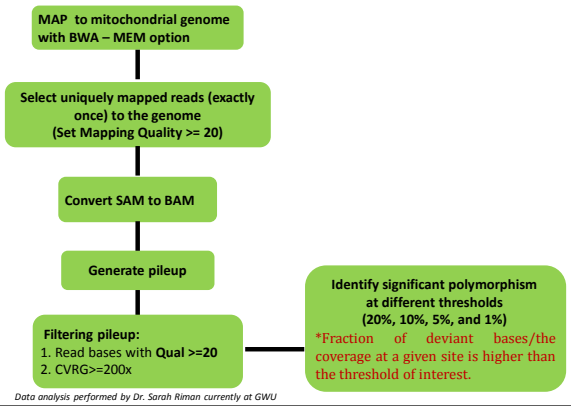


Variant calling thresholds were incrementally decreased

Variants observed on both PGM and MiSeq platforms were listed



Galaxy Pipeline for the Detection of Point Heteroplasmies



Data analysis performed by Dr. Sarah Riman currently at GWU

Detection of heteroplasmy as a function of allele calling thresholds

detected on both the MiSeq and PGM platforms

Threshold	2392 Component A	2392 Component B	2392-I
20%	64 C/T	7861 T/C	12071 T/C

64 C/T and 12071 T/C are listed in the current SRM certificates (detected by Sanger sequencing)

Data analysis performed by Dr. Sarah Riman currently at GWU

Detection of heteroplasmy as a function of allele calling thresholds

detected on both the MiSeq and PGM platforms

Threshold	2392 Component A	2392 Component B	2392-I
20%	64 C/T	7861 T/C	12071 T/C
10%	64 C/T	1393 G/A 7861 T/C	12071 T/C

64 C/T and 12071 T/C are listed in the current SRM certificates (detected by Sanger sequencing)

Data analysis performed by Dr. Sarah Riman currently at GWU

Detection of heteroplasmy as a function of allele calling thresholds

detected on both the MiSeq and PGM platforms

Threshold	2392 Component A	2392 Component B	2392-I
20%	64 C/T	7861 T/C	12071 T/C
10%	64 C/T	1393 G/A 7861 T/C	12071 T/C
5%	64 C/T	1393 G/A 7861 T/C	2445 T/C 5149 C/T 12071 T/C

64 C/T and 12071 T/C are listed in the current SRM certificates (detected by Sanger sequencing)

Data analysis performed by Dr. Sarah Riman currently at GWU

Detection of heteroplasmy as a function of allele calling thresholds

detected on both the MiSeq and PGM platforms

Threshold	2392 Component A	2392 Component B	2392-I
20%	64 C/T	7861 T/C	12071 T/C
10%	64 C/T	1393 G/A 7861 T/C	12071 T/C
5%	64 C/T	1393 G/A 7861 T/C	2445 T/C 5149 C/T 12071 T/C
1%	64 C/T	1393 G/A 3242 G/A 7412 C/T 7861 T/C	2445 T/C 5149 C/T 12071 T/C

64 C/T and 12071 T/C are listed in the current SRM certificates (detected by Sanger sequencing)

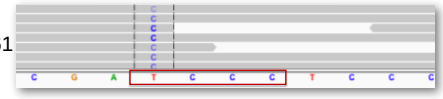
Data analysis performed by Dr. Sarah Riman currently at GWU

Detection of heteroplasmy as a function of sequencing platform and analysis pipeline

Platform & Analysis	2392 Comp A	2392 Comp B				2392-I		
	64 T	1393 A	3242 A	7412 T	7861 C	2445 C	5149 T	12071 C
PGM (CLC)	31%	15%	3%	3%	70%	7%	9%	50%
MiSeq (CLC)	32%	18%	4%	4%	88%	9%	6%	50%
PGM (Galaxy)	32%	15%	3%	ND	79%	8%	9%	50%
MiSeq (Galaxy)	31%	18%	5%	1.4%	89%	9%	7%	51%

The intent would be to include the lower level heteroplasmies as **informational values** in the updated certificate of analysis

Site 7861

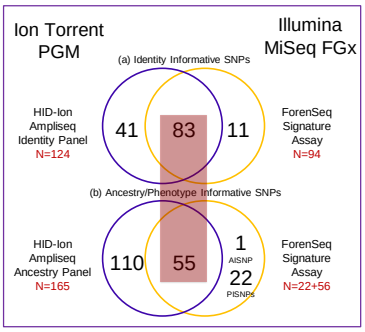


Data analysis performed by Dr. Sarah Riman currently at GWU

Forensic SNPs

- With NGS we now have a platform for SNP typing
  - Identity
  - Biogeographical ancestry
  - Phenotype
- *Ion torrent AmpliSeq HID and Ancestry panels (PGM)*
- *Illumina ForenSeq Kit SNPs (MiSeq/FGx)*
- *Thoughts about characterizing SNPs…we are certifying the measurement of a property (SNP allele) not the application of the data*

Forensic SNP Characterization



Forensic SNP Characterization
Testing nine candidate samples in replicate

- Two of the 138 overlapping SNPs indicate discordance
- rs7251928 (ancestry)
  - PGM = AC
  - MiSeq/FGx = CC
- rs4918664 (ancestry)
  - PGM = GG
  - MiSeq/FGx = AG

In the absence of overlapping markers we assess:
- Strand bias
- Coverage balance for heterozygous and homozygous genotypes
- Reproducibility (replicates)

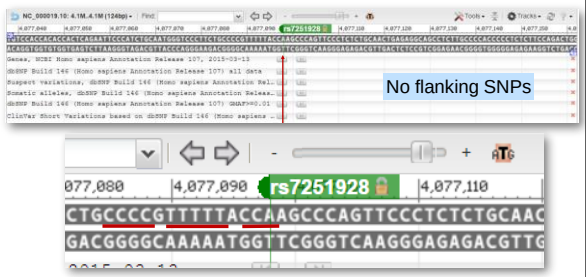
http://www.cstl.nist.gov/biotech/strbase/pub_pres/KieslerISFG2015poster.pdf

rs7251928 (A/C)

- MiSeq/FGx (CC)
  - 1000-2000x coverage
  - >99.9% C (no A calls)

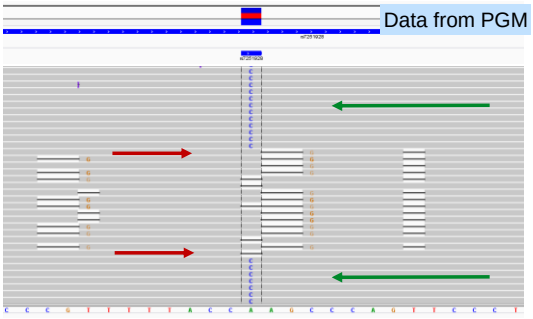
- PGM (AC)
  - 400-600x coverage
  - Strand bias (≈14%) – favors the 'negative strand'
  - Imbalanced allele coverage (14% A: 86% C)

rs7251928 (A/C)



Homopolymer regions around SNP site

rs7251928 (A/C)

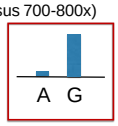


'Red' reads have issues going through SNP site

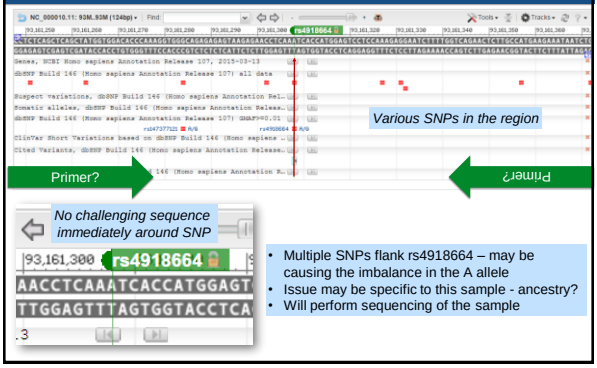
rs4918664 (A/G)

- MiSeq/FGx (AG)
  - Coverage 2000x
  - Balance = 0.68 (A coverage less than G)

- PGM (GG)
  - Strand bias (49-53%)
  - ≈ 5% A allele (a bit above background; QC filter flagged)
  - Lower coverage relative to other samples (300x versus 700-800x)
  - Other heterozygous samples are concordant
    - Issue may be specific to this one sample?



rs4918664 (A/G)



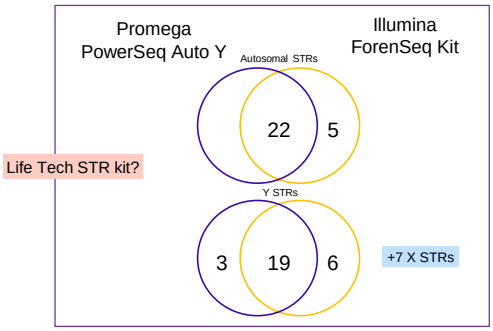
- Multiple SNPs flank rs4918664 – may be causing the imbalance in the A allele
- Issue may be specific to this sample - ancestry?
- Will perform sequencing of the sample

SRM 2391c

PCR-Based DNA Profiling Standard

- Data was recently updated (April 2015) - Sanger sequencing was performed for the **certified** STR loci
  - Sequencing motifs in certificate
  - https://www-s.nist.gov/srmors/view_cert.cfm?srm=2391C
- Future plans for NGS characterization of STR loci
  - Promega PowerSeq Auto/Y (STRait Razor)
  - Illumina ForenSeq (STRait Razor & Illumina UAS)
  - Other? (PGM)

Forensic STR Characterization



Sampling of SRM 2391c sequence data

Comp	Locus	Length-based	Sanger	Allele 1	Allele 2
A	D2S1338	18, 23	18, 23	$[TGCC]_6$ $[TTCC]_{12}$	$[TGCC]_7$ $[TTCC]_{13}$ **GTCC** $[TTCC]_2$
A	D2S441	10,10	10,10	$[TCTA]_{10}$	$[TCTA]_8$ **TCTG** [TCTA]
A	D8S1179	13,14	13,14	$[TCTA]_{13}$	$[TCTA]_2$ **TCTG** $[TCTA]_{11}$
B	D2S441	10,14	10,14	$[TCTA]_{10}$	$[TCTA]_{11}$ **TTTA** $[TCTA]_2$
B	D12S391	19,24	19,24	$[AGAT]_{12}$ $[AGAC]_6$ **AGAT**	$[AGAT]_{15}$ $[AGAC]_9$

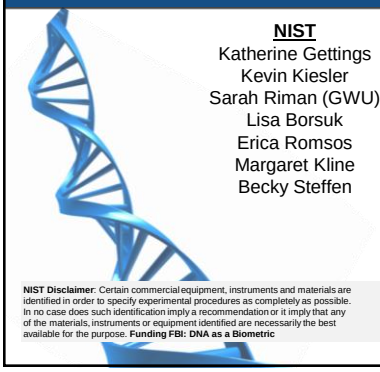
Examples of varying sequence motifs detected by sequencing

Future Strategy - Thoughts

- SRM 2391c typed the Promega and Illumina kits
  - No discordances observed from the certified CE/Sanger genotypes (in STR motifs only)
- SRM 2391d will be developed in the next two years
  - Characterize with multiple: platforms/assays/pipelines
  - Concordance to length based CE genotypes
  - Include flanking region sequence (how much?)
  - When will it be possible to replace Sanger sequencing of STRs for SRM work?

Acknowledgements

NIST
Katherine Gettings
Kevin Kiesler
Sarah Riman (GWU)
Lisa Borsuk
Erica Romsos
Margaret Kline
Becky Steffen



Contact Information
peter.vallone@nist.gov