

  
**Characterization of Reference Standards with  
Next-Generation Sequencing Platforms**  
  
 8<sup>th</sup> International Society of Applied Biological Sciences Conference  
 Split, Croatia, June 24-28, 2013  
  
 Peter M. Vallone, Ph.D.  
 Applied Genetics Group  
 US National Institute of Standards and Technology

---

---

---

---

---


---

---

---

**Overview of NIST**  
 National Institute of Standards and Technology

- National metrological institute (NMI)
- Founded in 1901
- ≈3000 employees across multiple campuses



Gaithersburg, Maryland Campus

- Maintains time measurement for the US (atomic clock)
- Four Nobel prize winners
- NIST supplies over 1,300 Standard Reference Materials (SRMs) for industry, academia, and government use in calibration of measurements

---

---

---

---


---

---

---

---

**Overview of Human Identity Project**



- **Past and current HID projects:**
- Develop and assess emerging technologies: miniSTRs, rapid PCR, new STR loci, mass spectrometry, SNPs, DNA mixture analysis, **high throughput sequencing (next-generation)**
- Training and workshops for the forensic DNA typing community (e.g. STRBase [www.cstl.nist.gov/strbase](http://www.cstl.nist.gov/strbase))
- **Reference materials (SRMs) to support forensic DNA typing measurements**

---

---

---

---

---

---


---

---

**NIST Standard Reference Materials**  
<http://www.nist.gov/srm/>

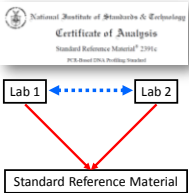
*Traceable standards to ensure accurate measurements in our nation's crime laboratories*

**Human Identity SRMs**  
 SRM 2391c – PCR-Based DNA Profiling  
 SRM 2392 & 2392-I – mitochondrial DNA  
 SRM 2395 – Y-STR DNA Profiling  
 SRM 2372 – Human DNA quantitation



SRM 2391c  
 Current price: \$626 USD

Genomic DNAs characterized for the expanded CODIS core loci and Y-STRs



Calibration with SRMs enables confidence in comparisons of results between laboratories

---

---

---

---

---

---

---

---

**Levels of Confidence**

Information contained in SRM certificate of analysis

- Certified** e.g. full Sanger sequencing to determine an STR allele
  - A NIST certified value is a value for which NIST has the **highest confidence** in its accuracy in that all known or suspected sources of bias have been investigated or taken into account.
- Reference** e.g. multiple PCR genotyping assays to determine an STR allele
  - A NIST Reference Value is a **best estimate** of the true value provided by NIST where all known or suspected sources of bias have not been fully investigated by NIST.
- Information** e.g. one PCR assay using ILS sizing to determine an STR allele
  - An information value is considered to be a value that will be of interest and use to the SRM user, but for which **insufficient information** is available to assess adequately the uncertainty associated with the value, or a value derived from a limited number of analyses.

May, W.E., Gillis, T.E., Parris, R., Beck, H.C.M., Fassett, J.D., Gettings, R.J., Greenberg, S.R., Guenther, F.A., Kramer, G., Macdonald, B.S., Wix, S.A., Definitions of Terms and Modes Used at NIST for Value-Assignment of Reference Materials for Chemical Measurements, NIST Special Publication 260-136 (2000); available at <http://ts.nist.gov/MeasurementServices/ReferenceMaterials/PUBLICATIONS.cfm>

---

---

---

---

---

---

---

---

**Characterization of the existing SRMs**

Current Status

- 2391c PCR Based DNA profiling standard**
  - 68 STR markers (51 autosomal + 17 Y chromosome)
  - STR repeat lengths (alleles) were certified using multiple (unique) PCR primer sets
  - Sanger sequencing was only performed for loci without multiple PCR primer sets (**only 10%**)
- 2392 & 2392-I Mitochondrial DNA sequencing standard**
  - Entire mtGenome (≈16,569 bp) was certified by Sanger sequencing
  - Multiple F/R strand coverage across the mtGenome

---

---

---

---

---

---

---

---

### Use of NGS for forensic applications

Highly-parallel/high-throughput next-generation sequencing technologies provide the ability to directly sequence forensically relevant targets  
Issues: sample input amounts, back compatibility, new workflows, cost, throughput, etc

- Whole mitochondrial genome analysis
  - Potential for improved sensitivity, mixture detection, multiplex sequencing of full mitochondrial genomes
  - Detection of minor SNP variants – heteroplasmy
- Going in depth into STR loci and beyond
  - STRs are useful for legacy (databases)
  - SNPs within STRs identify 'sub-alleles'
- Forensically relevant SNPs: newer human identity applications: biogeographical ancestry, externally visible traits, complex kinship, degraded samples, mixtures, low template, and other applications

---

---

---

---

---

---

---

---

### Initial Goals

- To characterize existing forensic SRMs with NGS
  - Further characterizes the materials with a new technique
  - Supports adoption of NGS in forensic community
  - 2391c: not all STR loci have full sequence information
  - 2392 and 2392-I: confirm Sanger data with a high coverage sequencing technology
  - Understand bias between NGS platforms: chemistry and bioinformatics
- Is there a need for a new material?
  - Forensic validation
  - Clinical validation

---

---

---

---

---

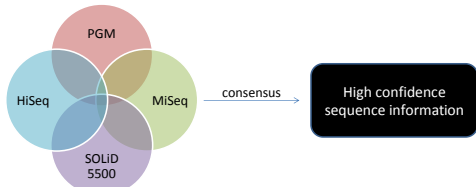
---

---

---

### Multiple NGS Platforms

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




---

---

---

---

---

---

---

---

## NIST SRM 2392 & 2392-I

- Mitochondrial DNA sequencing Standard Reference Materials
  - Characterized for mtDNA genome sequence composition
  - Reference used to validate measurement techniques
  - Recommended by FBI as positive control for sequencing labs
- SRM 2392
  - Contains 3 components (extracted DNA)
    - 2392 A – From cell line CHR
    - 2392 B – From cell line 9947A
    - 2392 C – Cloned region of heteroplasmy
- SRM 2392-I
  - From cell line HL-60




---

---

---

---

---

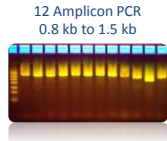
---

---

---

## Sequencing studies were performed on four NGS platforms

- **Ion Torrent PGM**
  - Edge Biosystems (outsourced)
  - Experiments performed at NIST
- **Illumina HiSeq 2000**
  - Beckman-Coulter Genomics (outsourced)
- **Illumina MiSeq**
  - Edge Biosystems (outsourced)
- **SOLiD 5500**
  - Experiments performed at NIST



Illumina MiSeq and HiSeq platforms will be online at NIST by the end of this summer

---

---

---

---

---

---

---

---

## Data Processing, Alignment, and Variant Calling

	Ion Torrent PGM	Illumina MiSeq	Illumina HiSeq	SOLID 5500
<b>Signal Processing</b> Output: <b>FASTQ</b>	Torrent Server	MiSeq Reporter	HiSeq Control	LifeScope
<b>Read Mapping</b> Output: <b>BAM</b>	Torrent Server	Novoalign	BWA	LifeScope
<b>Variant Calling</b> Output: <b>VCF</b>	Torrent Server	GATK	GATK	GATK

Abbreviations:  
 FASTQ – Unaligned reads in text format with quality scores  
 BAM – Binary Alignment Map (Aligned reads)  
 VCF – Variant Call File  
 BWA – Burrows Wheeler Aligner  
 GATK – Genome Analysis Tool Kit

---

---

---

---

---

---

---

---

### Sequence Coverage Summary

Experiment	Average Read Depth (MQ20*)	Experiment Design
EdgeBio PGM	280 x	Seven mtGenomes + spike-in controls**
NIST PGM Run 1	6,500 x	Three mtGenomes
NIST PGM Run 2	9,000 x	Three mtGenomes
Illumina MiSeq	49,000 x	Seven mtGenomes
Illumina HiSeq	41,000 x	Seven mtGenomes + spike-in controls**
NIST SOLID	29,000 x	Seven mtGenomes + spike-in controls**

\* MQ20 = reads with aligned quality score of 20 or above = less than 1 error per 100 bases

\*\*Spike-in control was NIST SRM 2374: DNA Sequence Library for External RNA Controls

---

---

---

---

---

---

---

---

---

---

---

---

### False Positives and False Negatives

Using platform specific informatics pipeline

		PGM 1	PGM 2	PGM 3	HiSeq	MiSeq	5500
9947A	FP	1	5	3	21	9	11
	FN	3	4	3	3	3	3
CHR	FP	2	6	10	21	9	10
	FN	3	5	4	3	3	4
HL-60	FP	1	8	8	20	9	8
	FN	1	2	1	1	1	1
Avg Coverage		280	6,500	9,000	49,000	41,000	29,000

Calls made to the rCRS  
On average 99.94 % agreement with Sanger sequencing

---

---

---

---

---

---

---

---

---

---

---

---

### False Positives and False Negatives

Using platform specific informatics pipeline

		PGM 1	PGM 2	PGM 3	HiSeq	MiSeq	5500
9947A	FP	1	5	3	21	9	11
	FN	3	4	3	3	3	3
CHR	FP	2	6	10	21	9	10
	FN	3	5	4	3	3	4
HL-60	FP	1	8	8	20	9	8
	FN	1	2	1	1	1	1
Avg Coverage		280	6,500	9,000	49,000	41,000	29,000

False negatives were concentrated in C stretch regions of the genome  
The FN sites 13,759 and 5,228 were due to low coverage

9947A (FN) = 309.1, 309.2, 315.1, 13,759  
CHR (FN) = 309.1, 315.1, 16193.1, 16183, 16189  
HL-60 (FN) = 315.1, 5,228

---

---

---

---

---

---

---

---

---

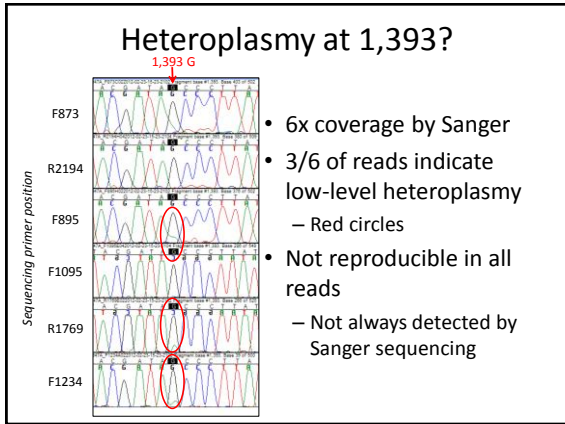
---

---

---








---

---

---

---

---

---

---

---

---

---

### Heteroplasmy detected by NGS at Site 1,393

- Agreement across platforms (**high confidence**)
- $\approx 17.6\% (\pm 2.6\%)$  minor component "A"

Experiment	Reference "G"	Variant "A"	Coverage
EdgeBio PGM	77.3%	22.7%	97 x
NIST PGM Run 1	82.5%	17.5%	2940 x
NIST PGM Run 2	83.4%	16.6%	3275 x
Illumina MiSeq	83.7%	16.3%	26,234 x
Illumina HiSeq	84.4%	15.6%	62,186 x
NIST SOLiD	82.5%	16.9%	24,226 x

Site 1,393 also confirmed by Niels Morling's lab using 454 technology (Martin Mikkelsen)

---

---

---

---

---

---

---

---

---

---

- ### 'Challenging' Samples
- Mitochondrial haplogroups
- C1-3-02
    - 48 differences from rCRS, del-249, 290, 291, 523, 524
  - L1b-2-03
    - 87 differences from rCRS, del 523, 524, Cins-573.1-573.4
  - B2-1-07
    - 46 differences from rCRS del-8281-8289
  - L0a-1-03
    - 90 differences from rCRS, del-523, 524
- Selected from NIST US population samples

---

---

---

---

---

---

---

---

---

---



## False Positives and False Negatives

Using platform specific informatics pipeline

		PGM	HiSeq	MiSeq	5500
C1-3-02	FP	1	29	17	18
→	FN	8	8	7	8
L1b-2-03	FP	1	25	15	11
→	FN	10	9	9	11
B2-1-07	FP	4	25	9	16
→	FN	17	14	15	14
L0a-1-03	FP	2	21	10	8
→	FN	6	5	4	4

---

---

---

---

---

---

---

---

## False Negatives

common across the 4 platforms

- C1-3-02
  - 48 differences from rCRS, del-249, 290, 291, 523, 524
  - FN: 249, 290, 291, 309.1, 315.1, 523, 524
- L1b-2-03
  - 87 differences from rCRS, del-523, 524, Cins-573.1-573.4
  - FN: 309.1, 315.1, 523, 524, 573.1-4, 14560
- B2-1-07
  - 46 differences from rCRS del-8281-8289
  - FN: 309.1, 309.2, 315.1, 8281-8289, 16182, 16183, 16189
- L0a-1-03
  - 90 differences from rCRS, del-523,524
  - FN: 315.1, 523, 524

---

---

---

---

---

---

---

---

## Summary

- The consensus data from the four NGS platforms for the mitochondrial SRMs agree with Sanger sequencing data
  - G/A heteroplasmy at 1,393 confirmed
  - C insertions and deletions are issues (assemblers/variant callers)
  - The majority of false positives are of low abundance and not reproducible across platforms
- Continuing work
  - Experiments for setting a variant calling threshold
  - Evaluate a three amplicon approach for mitochondrial DNA enrichment
  - Sequence the mitoSRMs on the PacificBiosciences platform (Collaboration with Children's National Medical Center)
  - Benefit from a standardized (forensic) informatics pipeline (CLC bio software, NextGENe)
- 2391c characterization
  - Sanger and NGS sequencing of STR alleles
  - Beta test: LifeTech SNP panel, Illumina assays this fall

Clinical reference material – 'Genome in a Bottle'

---

---

---

---

---

---

---

---

### Genome in a Bottle Consortium

a NIST-hosted Public-Private-Academic partnership to develop infrastructure that enables **clinical application** of human genome sequencing

- genomic DNA reference materials will be developed to characterize performance of genome sequencing
  - materials will be certified for their variants against a reference sequence, with confidence estimates
- consortium will develop sequencing performance metrics that will enable confidence in results
  - enable regulatory oversight of clinical applications
  - genetic diseases, cancer, pharmacogenomics, transplant typing...

Genome in a Bottle Consortium <http://genomeinabottle.org>

---

---

---

---

---

---

---

---

### Integration of Data to Form "Gold Standard" Genotype Calls

Candidate variants	Find all possible variant sites
Concordant variants	Find highly confident sites across multiple datasets
Find characteristics of bias	Identify sites with atypical characteristics signifying sequencing, mapping, or alignment bias
Arbitration	For each site, remove datasets with decreasingly atypical characteristics until all datasets agree
Confidence Level	Even if all datasets agree, identify them as uncertain if few have typical characteristics, or if they fall in known segmental duplications or long repeats

30

---

---

---

---

---

---

---

---

### Thank you for your attention!

Questions?  
 peter.vallone@nist.gov  
 1-301-975-4872

Acknowledgements  
**Kevin Kiesler** and Mike Coble

Multiplexed Biomolecular Science Group  
 Jenny McDaniel, Justin Zook and Marc Salit

Outside funding agencies:  
 FBI - Evaluation of Forensic DNA Typing as a Biometric Tool  
 NIJ - Interagency Agreement with the Office of Law Enforcement Standards

---

---

---

---

---

---

---

---