

# Capabilities of Next-Generation Sequencing Instrumentation for Mitochondrial DNA Whole Genome Analysis

Presentation to  
Mid-Atlantic Association of Forensic Scientists Meeting  
May 8, 2013

Kevin Kiesler  
National Institute of Standards and Technology  
Applied Genetics Group

# Outline

- Mitochondrial DNA background
- NIST SRM for mtDNA sequencing
- Next-generation sequencing technology
- Experimental methods
- Experimental results



# NIST SRM 2392 & 2392-1

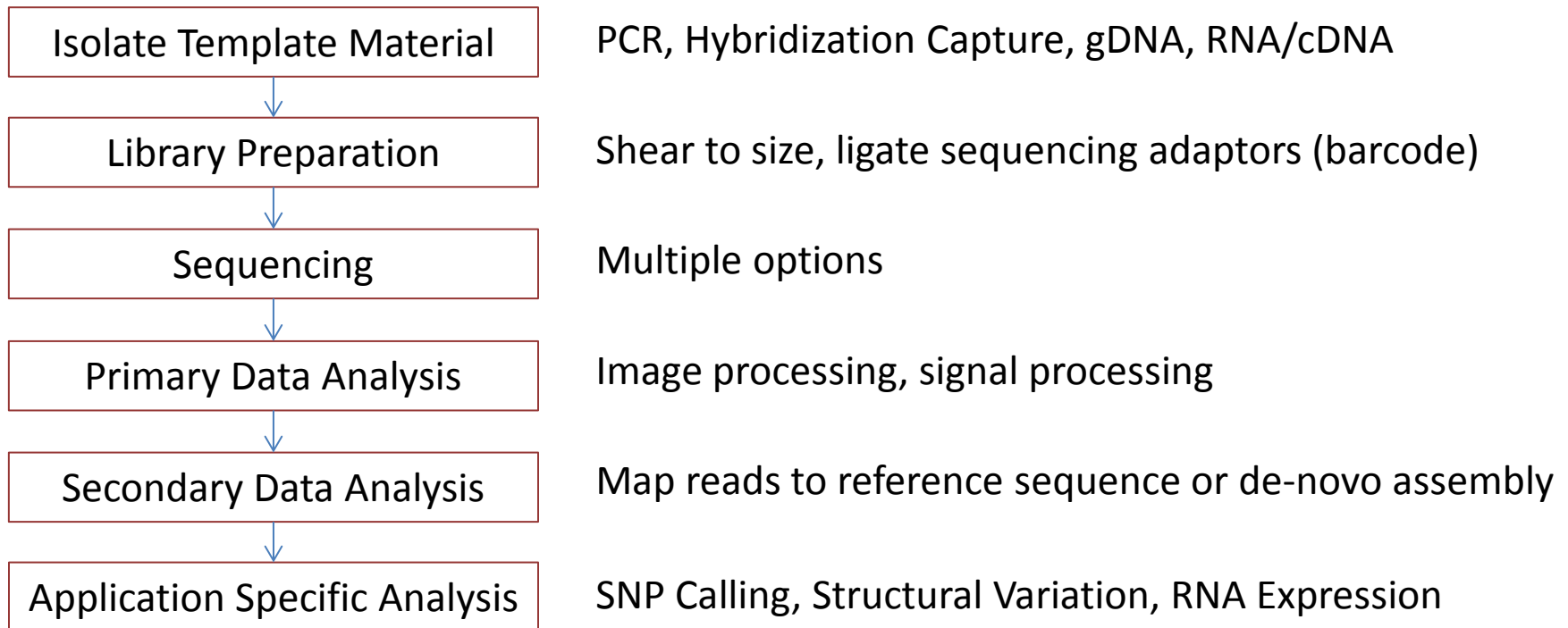
- Mitochondrial DNA sequencing Standard Reference Materials
  - Characterized for mtDNA genome sequence composition
  - Reference used to validate measurement techniques
  - Recommended by FBI as positive control for sequencing labs
- SRM 2392
  - Contains 3 components (extracted DNA)
    - 2392 A – From cell line CHR
    - 2392 B – From cell line 9947A
    - 2392 C – Cloned region of heteroplasmy
- SRM 2392-1
  - From cell line HL-60



# New Technology for Sequencing

- SRMs were initially characterized with Sanger sequencing
  - Levin et al. NIST Special Publication 260-155 (2003)
  - <http://www.nist.gov/srm/upload/sp260-155.pdf>
- Why use next generation sequencing (NGS)?
  - Whole mitochondrial genome analysis = **more information**
  - Potential for improved sensitivity
    - Detection of minor SNP variants - heteroplasmy
  - **Confirm SRM sequence** with an orthogonal technique
- Initial approach for NIST experiments
  - Sequence on multiple NGS platforms
  - Understand differences between platforms
  - Gain practical experience in library preparation, sequence data generation, and assembly/variant calling

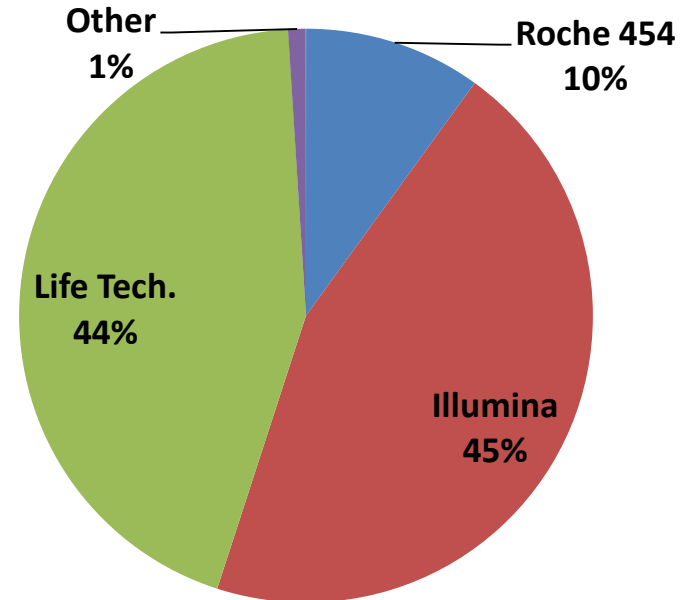
# Next Generation Sequencing Process



# Four Common NGS Platforms

- Ion Torrent (Life Tech.)
  - Semiconductor sequencing
- SOLiD (Life Tech.)
  - Sequencing by ligation
- Illumina
  - Sequencing by synthesis
- Roche 454
  - Pyrosequencing

Market Share for Desktop Instruments  
(Approximate)



Source: DeciBio NGS Report (March 2013)  
[http://www.decibio.com/NGS\\_PR](http://www.decibio.com/NGS_PR)

# Platform Output and Cost

Values estimated from manufacturer's specifications

	Output (Gb)	Run Time (Hours)	Cost Per Gb
Ion Torrent PGM ★	2	4.5	\$
Ion Torrent Proton ★	10	4	\$\$
Illumina MiSeq ★	8	48	\$
Illumina Genome Analyzer	95	336	\$\$
Illumina HiSeq 2000/2500	600	264	\$\$\$
Life Technologies SOLiD	320	168	\$\$\$
Roche 454 GS Junior ★	0.035	10	\$\$
Roche 454 GS FLX	0.7	23	\$\$\$
Sanger sequencing (ABI 3730xl)	0.00005	2	\$\$\$\$

★ “Benchtop sequencer” – lower instrument cost, smaller size, shorter run time



# Next Generation Sequencing Experiments

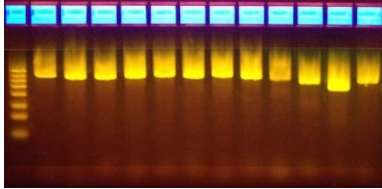
# Approach

Pilot studies were performed on 4 platforms

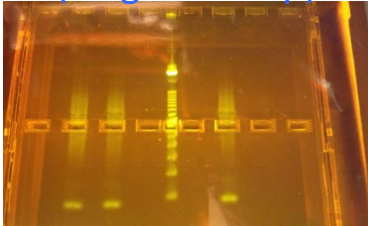
- **Ion Torrent PGM**
  - Edge Biosystems (outsourced)
  - Instrument at NIST
- **Illumina HiSeq 2000**
  - Beckman-Coulter Genomics (outsourced)
- **Illumina MiSeq**
  - Edge Biosystems (outsourced)
- **SOLiD 5500**
  - Instrument at NIST

# Ion Torrent PGM Workflow

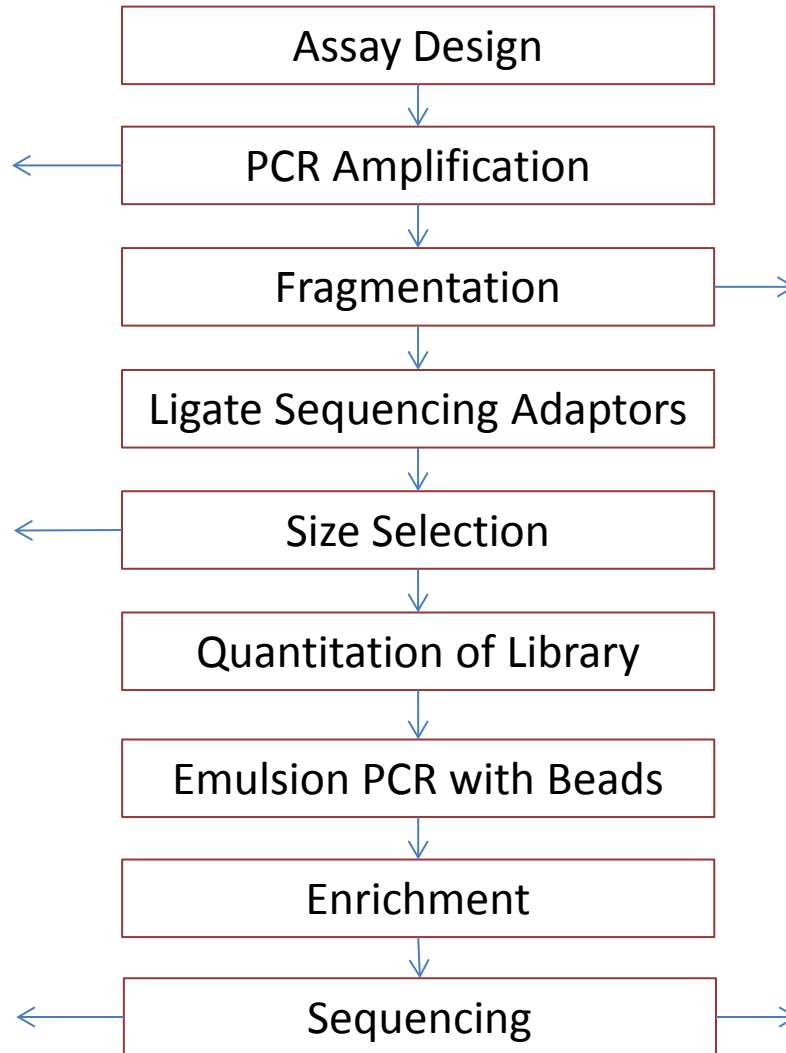
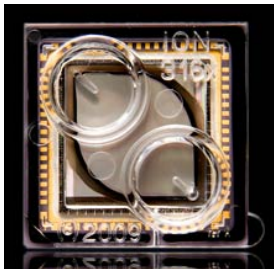
12 Amplicon PCR  
0.8 kb to 1.5 kb



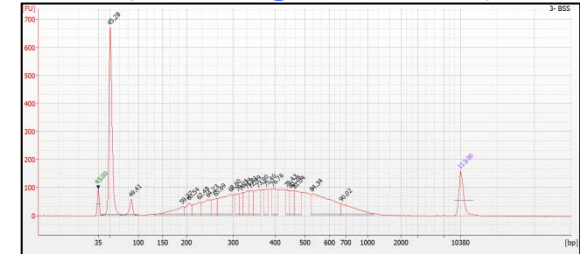
Size Selection Gel  
(Target 250 bp)



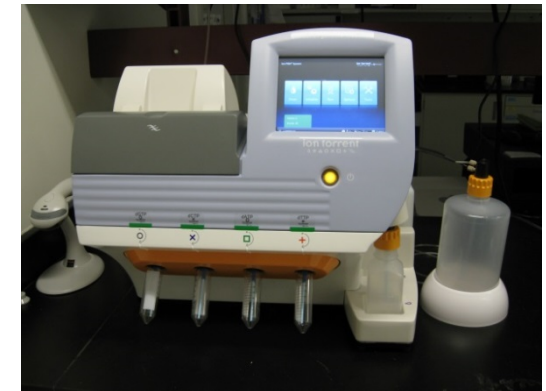
Ion 316 Chip  
(6.1 Million Wells)



Fragmentation  
(Size range 100-1000)



Ion Torrent PGM Instrument



# 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

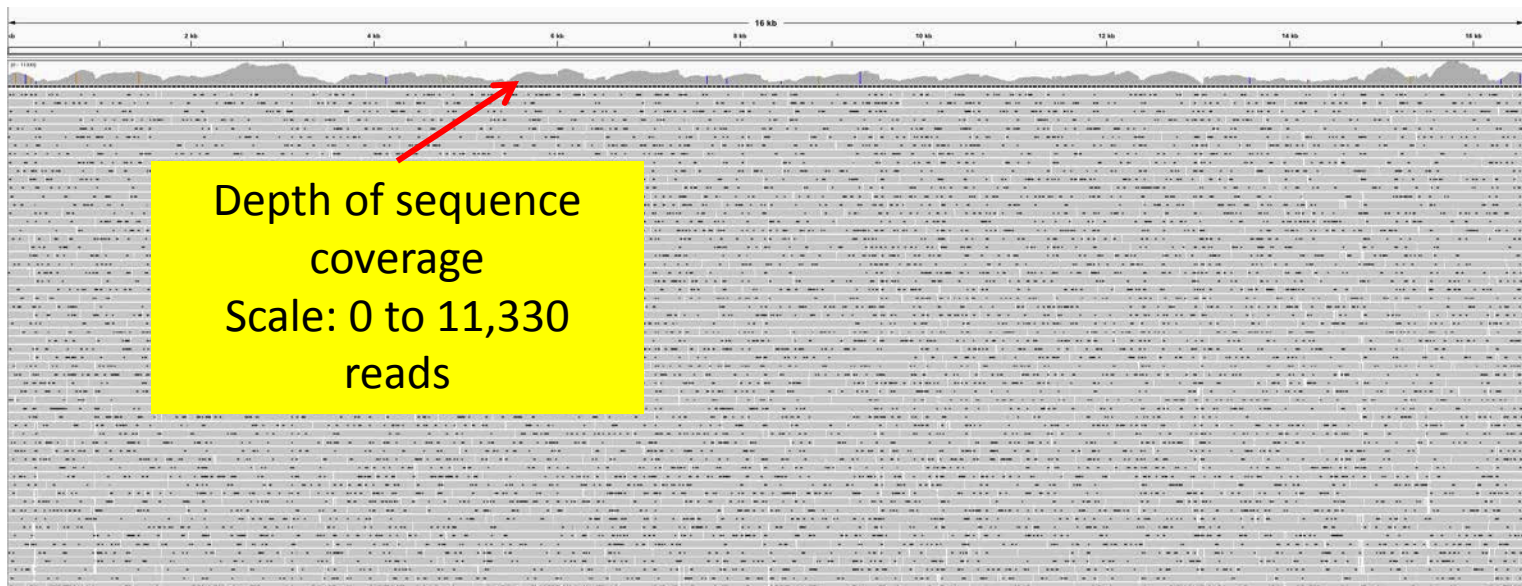
Experiment	Average Read Depth (AQ20*)	Experiment Design
EdgeBio PGM	280 x	Seven mtGenomes + spike-in controls**
NIST PGM Run 1	6,500 x	Three mtGenomes
NIST PGM Run 2	9,000 x	Three mtGenomes
Illumina MiSeq	49,000 x	Seven mtGenomes
Illumina HiSeq	41,000 x	Seven mtGenomes + spike-in controls**
NIST SOLiD	29,000 x	Seven mtGenomes + spike-in controls**

\* AQ20 = reads with aligned quality score of 20 or above  
= less than 1 error per 100 bases

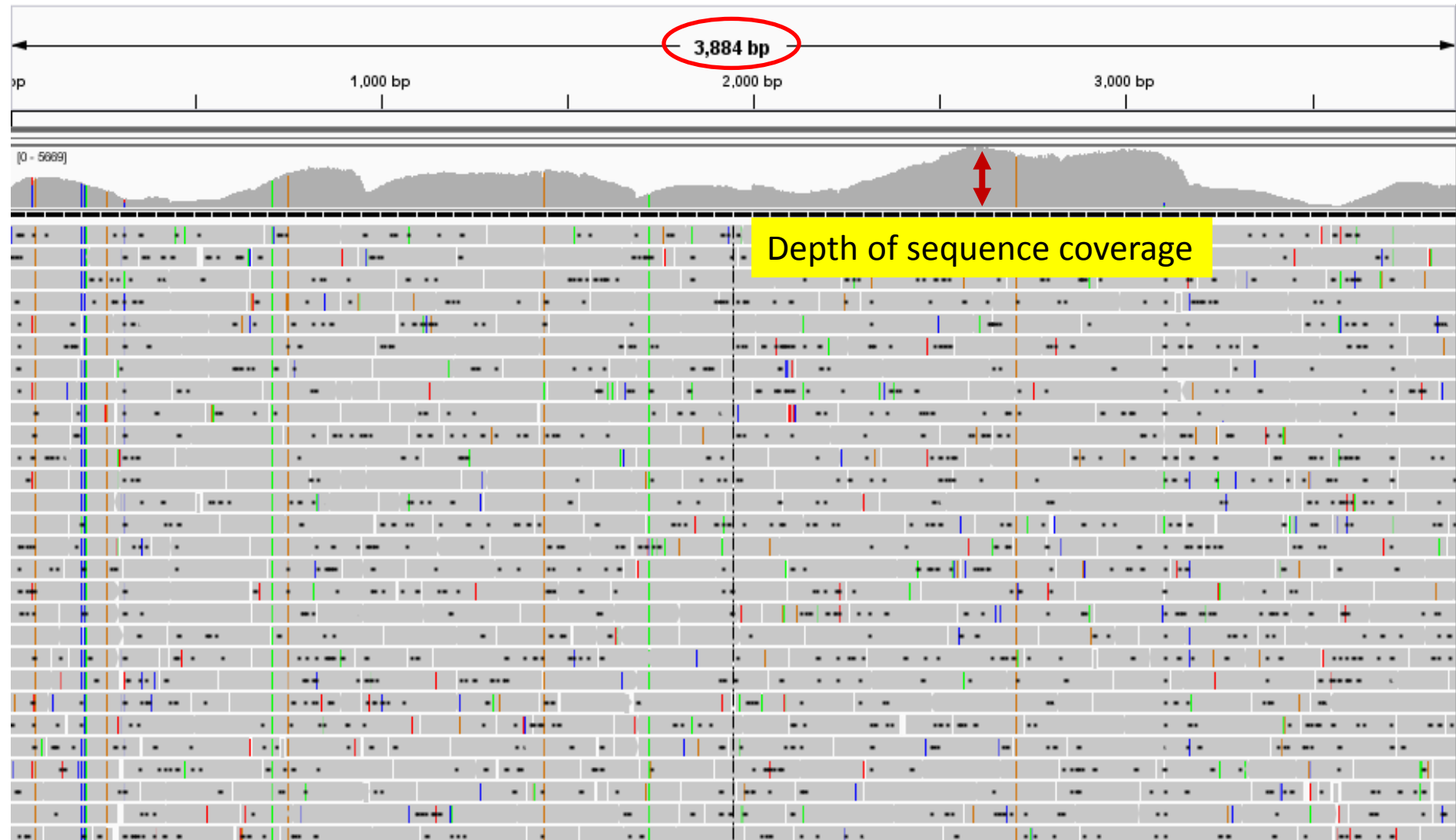
\*\*Spike-in control was NIST SRM 2374: DNA Sequence Library for External RNA Controls

# Broad Overview of Sequence Data

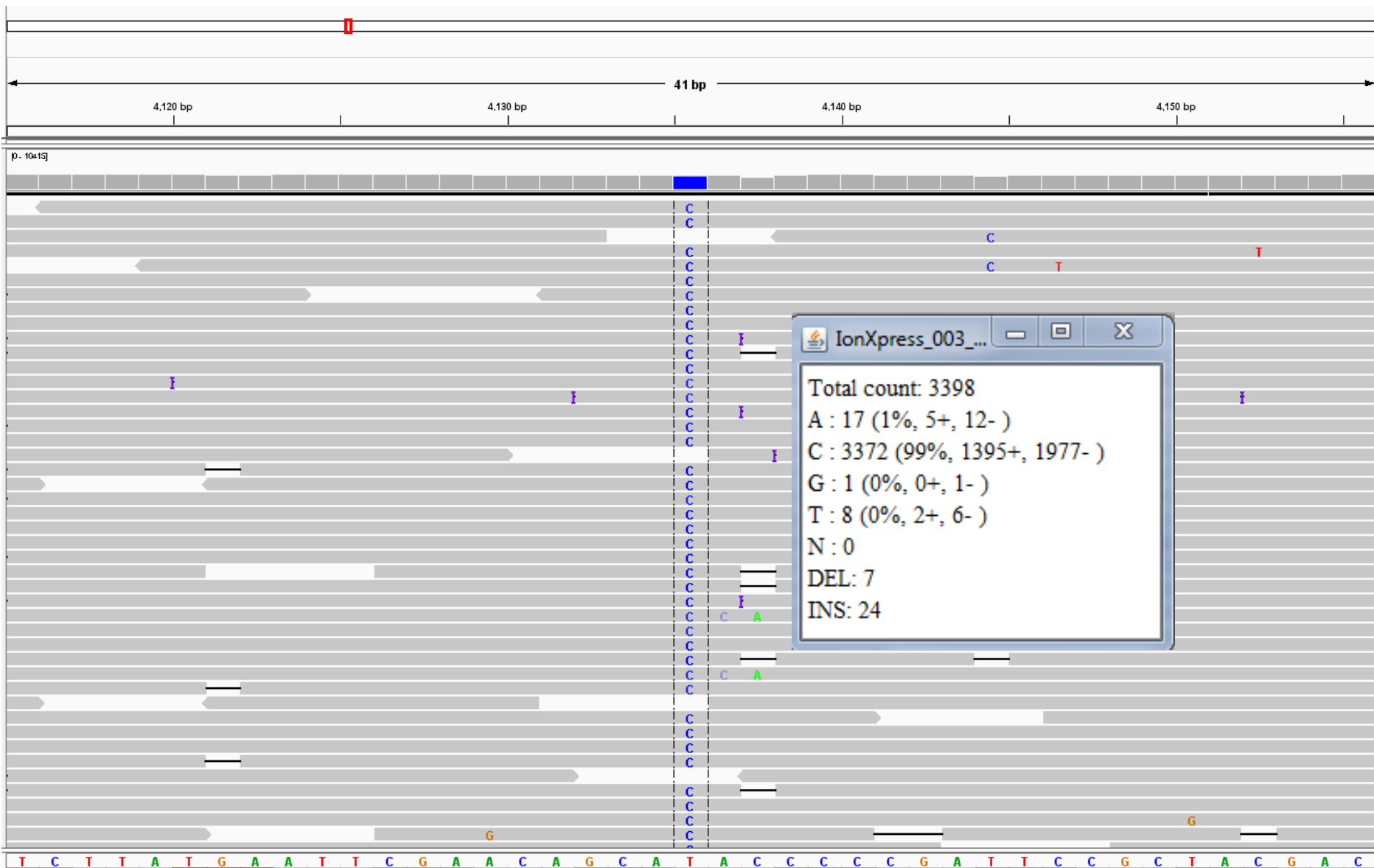
- Integrative Genomics Viewer (IGV)
  - Displays aligned sequences
    - Accepts BAM files
    - Sequence and coverage information displayed
  - Freely available from Broad Institute
    - Joint MIT/Harvard biomedical institute



# Sequence Data - Zoomed In



# Sequence Data – Individual Bases





# Variant Calls – Concordance

- Differences versus revised Cambridge Reference Sequence (rCRS) positions are reported
  - 16,569 bases
  - Only sites which differ are reported
  - All other sites are in agreement with reference
- Nucleotide position, reference & variant genotypes
- Other outputs: coverage statistics, quality metrics, p-values, etc.

Position	Type	Zygoty	Ref	Variant	Coverage	Ref Cov	Var Cov
93	SNP	Hom	A	G	4628	2	4626
195	SNP	Hom	T	C	3568	35	3533
214	SNP	Hom	A	G	3106	58	3046
263	SNP	Hom	A	G	2472	4	2460
750	SNP	Hom	A	G	3957	30	3919
1438	SNP	Hom	A	G	4826	57	4766
3106	SNP	Het	C	A	345	262	78
4135	SNP	Hom	T	C	2645	0	2642
4769	SNP	Hom	A	G	2620	10	2603
7645	SNP	Hom	T	C	3029	12	3016
7861	SNP	Het	T	C	1241	278	959
8448	SNP	Hom	T	C	1202	3	1187
8860	SNP	Hom	A	G	2991	5	2980
9315	SNP	Hom	T	C	4469	12	4456
13058	SNP	Het	C	A	198	131	58
13572	SNP	Hom	T	C	2528	8	2519
13759	SNP	Hom	G	A	814	2	758
14199	SNP	Het	T	G	1745	912	833
15326	SNP	Hom	A	G	3038	11	3023
16311	SNP	Hom	T	C	2097	10	2084
16519	SNP	Hom	T	C	3601	8	3582

Table output from Life Technologies  
“Torrent Server” software

# Variant Calls – Concordance

## SRM 2392 Component B (9947A)

Nucleotide Position	rCRS Reference Sequence	SRM 2392 Component B Sanger Call	EdgeBio PGM	NIST PGM run 1	NIST PGM run 2	EdgeBio Illumina MiSeq	Beckman Genomics Illumina HiSeq	NIST SOLiD
93	A	G	G	G	G	G	G	G
195	T	C	C	C	C	C	C	C
214	A	G	G	G	G	G	G	G
263	A	G	G	G	G	G	G	G
309.1	:	C						
309.2	:	C						
315.1	:	C						
750	A	G	G	G	G	G	G	G
1393	G	G	G/A	G/A	G/A	G/A	G/A	G/A
1438	A	G	G	G	G	G	G	G
4135	T	C	C	C	C	C	C	C
4769	A	G	G	G	G	G	G	G
7645	T	C	C	C	C	C	C	C
7861	T	C	C	C	C	C	C	C
8448	T	C	C	C	C	C	C	C
8860	A	G	G	G	G	G	G	G
9315	T	C	C	C	C	C	C	C
13572	T	C	C	C	C	C	C	C
13759	G	A	A		A	A	A	A
15326	A	G	G	G	G	G	G	G
16311	T	C	C	C	C	C	C	C
16519	T	C	C	C	C	C	C	C

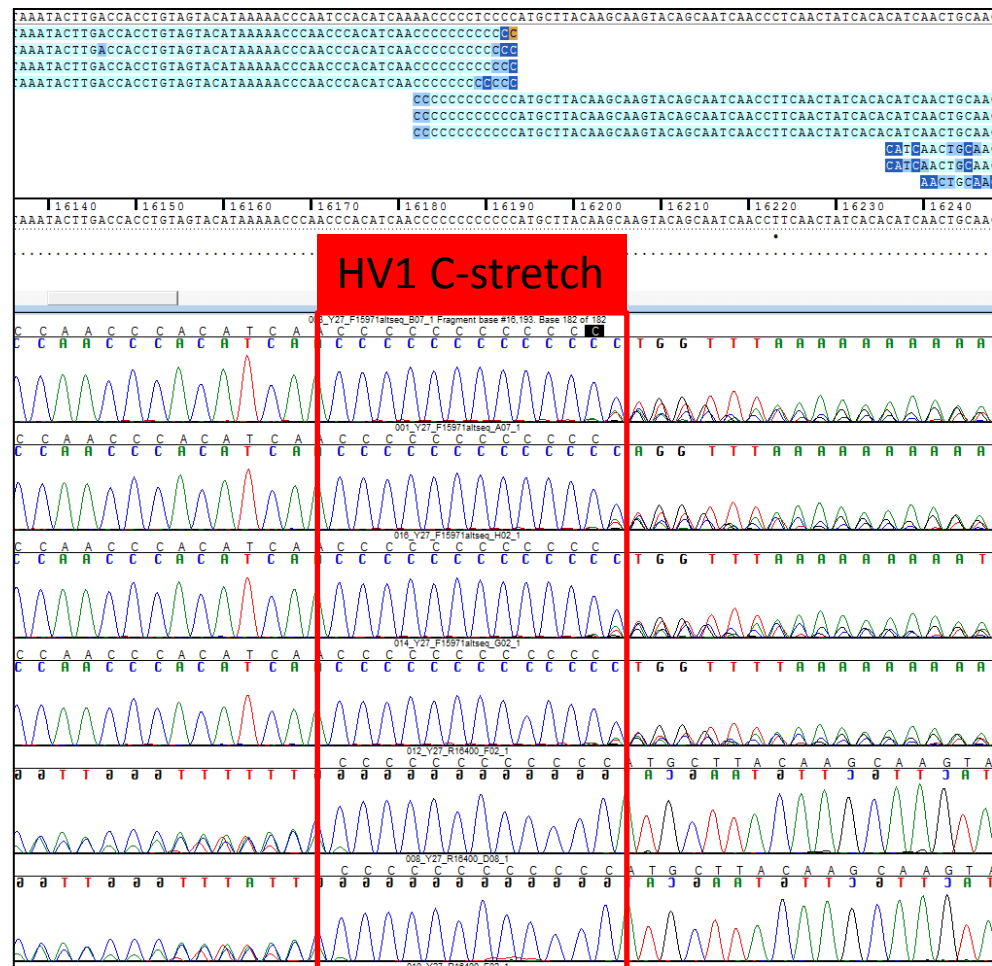
# Insertions in HV2 C-stretch not correctly called according to forensic nomenclature rules

Nucleotide Position	rCRS Reference Sequence	SRM 2392 Component B Sanger Call	EdgeBio PGM	NIST PGM run 1	NIST PGM run 2	EdgeBio Illumina MiSeq	Beckman Genomics Illumina HiSeq	NIST SOLiD
93	A	G	G	G	G	G	G	G
195	T	C	C	C	C	C	C	C
214	A	G	G	G	G	G	G	G
263	A	G	G	G	G	G	G	G
309.1	:	C						
309.2	:	C						
315.1	:	C						
750	A	G	G	G	G	G	G	G
1393	G	G	G/A	G/A	G/A	G/A	G/A	G/A
1438	A	G	G	G	G	G	G	G
4135	T	C	C	C	C	C	C	C
4769	A	G	G	G	G	G	G	G
7645	T	C	C	C	C	C	C	C
7861	T	C	C	C	C	C	C	C
8448	T	C	C	C	C	C	C	C
8860	A	G	G	G	G	G	G	G
9315	T	C	C	C	C	C	C	C
13572	T	C	C	C	C	C	C	C
13759	G	A	A		A	A	A	A
15326	A	G	G	G	G	G	G	G
16311	T	C	C	C	C	C	C	C
16519	T	C	C	C	C	C	C	C

# Control Region Homopolymers

- Two regions of poly C interrupted with one T in the middle, termed “C-stretch”
  - Difficult to sequence through with Sanger chemistry
- Challenge to align
  - There can be multiple length variants (heteroplasmy) within one sample
  - Strict nomenclature rules are used
- NGS variant calling software with forensic mtDNA nomenclature rules not yet available

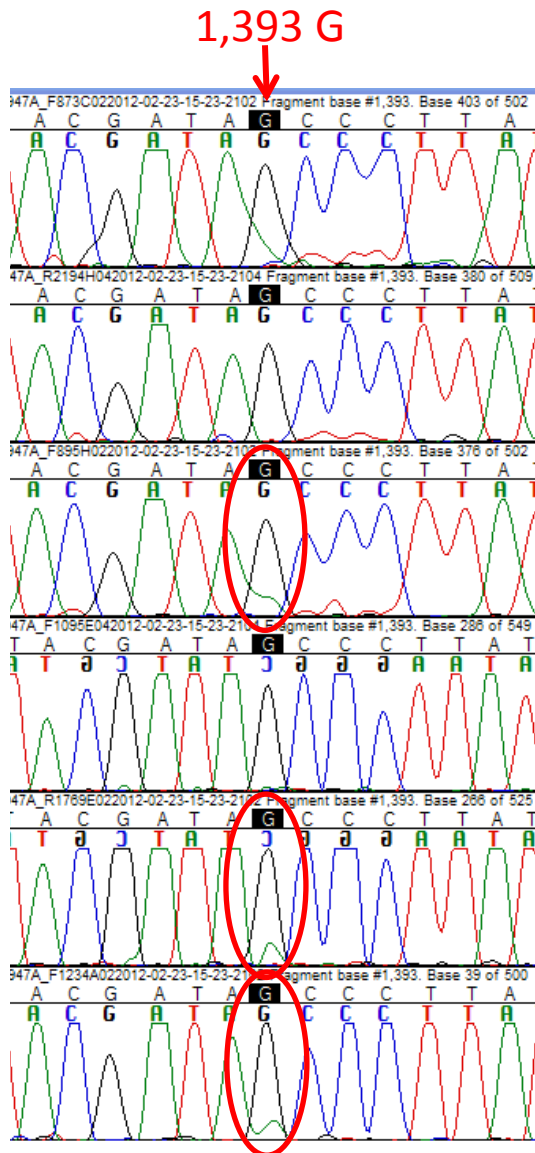
Reference sequence  
HV1 = CCCCCCTCCCC  
HV2 = CCCCCCCCTCCCCC



# Heteroplasmy at Position 1,393

Nucleotide Position	rCRS Reference Sequence	SRM 2392 Component B Sanger Call	EdgeBio PGM	NIST PGM run 1	NIST PGM run 2	EdgeBio Illumina MiSeq	Beckman Genomics Illumina HiSeq	NIST SOLiD
93	A	G	G	G	G	G	G	G
195	T	C	C	C	C	C	C	C
214	A	G	G	G	G	G	G	G
263	A	G	G	G	G	G	G	G
309.1	:	C						
309.2	:	C						
315.1	:	C						
750	A	G	G	G	G	G	G	G
1393	G	G	G/A	G/A	G/A	G/A	G/A	G/A
1438	A	G	G	G	G	G	G	G
4135	T	C	C	C	C	C	C	C
4769	A	G	G	G	G	G	G	G
7645	T	C	C	C	C	C	C	C
7861	T	C	C	C	C	C	C	C
8448	T	C	C	C	C	C	C	C
8860	A	G	G	G	G	G	G	G
9315	T	C	C	C	C	C	C	C
13572	T	C	C	C	C	C	C	C
13759	G	A	A		A	A	A	A
15326	A	G	G	G	G	G	G	G
16311	T	C	C	C	C	C	C	C
16519	T	C	C	C	C	C	C	C

# Heteroplasmy at 1,393?



- 3/6 of Sanger reads show possible low-level heteroplasmy
  - Red circles
- Not reproducible in all reads
  - Not always detected by Sanger sequencing

# Heteroplasmy detected by NGS at Site 1,393

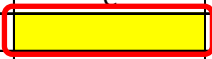
- Agreement across platforms
  - $\approx 17.6\%$  ( $\pm 2.6\%$ ) minor component “A”

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

# 'Missed Call' at Position 13,759

Nucleotide Position	rCRS Reference Sequence	SRM 2392 Component B Sanger Call	EdgeBio PGM	NIST PGM run 1	NIST PGM run 2	EdgeBio Illumina MiSeq	Beckman Genomics Illumina HiSeq	NIST SOLiD
93	A	G	G	G	G	G	G	G
195	T	C	C	C	C	C	C	C
214	A	G	G	G	G	G	G	G
263	A	G	G	G	G	G	G	G
309.1	:							
309.2	:							
315.1	:							
750	A							G
1393	C							G/A
1438	A							G
4135	T							C
4769	A							G
7645	T	C	C	C	C	C	C	C
7861	T	C	C	C	C	C	C	C
8448	T	C	C	C	C	C	C	C
8860	A	G	G	G	G	G	G	G
9315	T	C	C	C	C	C	C	C
13572	T	C	C	C	C	C	C	C
13759	G	A	A		A	A	A	A
15326	A	G	G	G	G	G	G	G
16311	T	C	C	C	C	C	C	C
16519	T	C	C	C	C	C	C	C

On closer inspection of NIST PGM run 1:  
 1155 reads at this position were "A" (correct)  
 Only reverse strand reads were generated  
**Strand bias resulted in 'missed call'**





# Variant Calls – False Positives

- Variants denoted by nucleotide position and base call
- Percentage of reads associated with variant calls shown
- Ion Torrent PGM variants called with Torrent Suite software
- Illumina and SOLiD variants called with GATK Unified Genotyper

PGM	%	MiSeq	%	HiSeq	%	SOLiD	%
3106 A	38.8%	302 INS C		302 INS C		301 INS C	
5744 C	3.9%	308 INS C		309 INS C		302 INS C	
7861 DEL	21.3%	310 INS C		310 INS C		309 INS C	
11512 A	6.6%	514 DEL		310 INS C		310 INS C	
13045 INS C	25.8%	515 DEL		360 T	0.4%	347 A	1.4%
13058 A	19.3%	515 G	0.9%	1992 T	1.1%	4722 G	0.4%
		3106 A	0.9%	3103 T	0.6%	6482 A	0.3%
		12418 C	0.4%	3104 C	1.2%	15284 C	0.1%
				3105 T	2.1%		
				4796 T	1.0%		
				6419 C	26.1%		
				8163 T	2.1%		
				9753 T	2.4%		
				14188 T	0.4%		
				15259 T	1.9%		
				15877 T	0.5%		
F.P. = 6		F.P. = 8		F.P. = 16		F.P. = 8	

# Variant Calls – False Positives

- Many false positives are low-level
  - 0.1% to 2.4% of reads
  - Only in GATK variant calls
- These can be filtered out computationally
  - Evaluation of thresholds is recommended

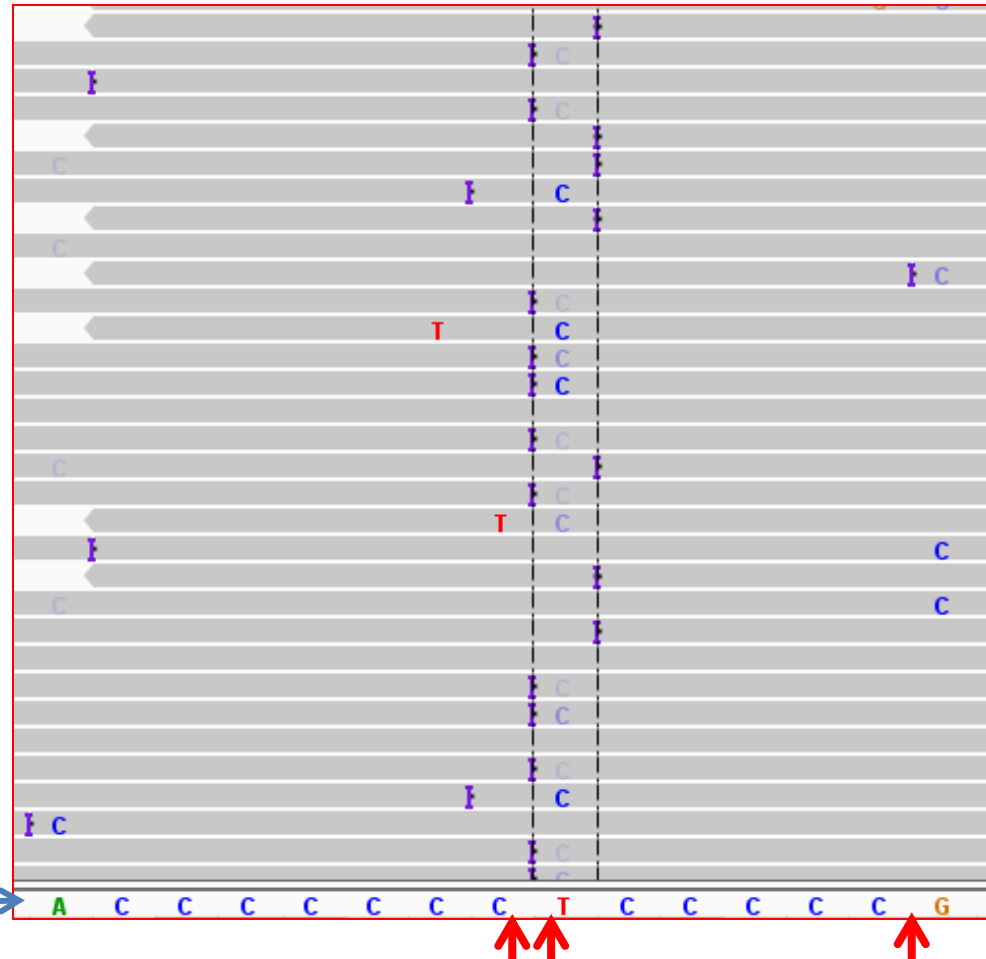
PGM	%	MiSeq	%	HiSeq	%	SOLiD	%
3106 A	38.8%	302 INS C		302 INS C		301 INS C	
5744 C	3.9%	308 INS C		309 INS C		302 INS C	
7861 DEL	21.3%	310 INS C		310 INS C		309 INS C	
11512 A	6.6%	514 DEL		310 INS C		310 INS C	
13045 INS C	25.8%	515 DEL		360 T	0.4%	347 A	1.4%
13058 A	19.3%	515 G	0.9%	1992 T	1.1%	4722 G	0.4%
		3106 A	0.9%	3103 T	0.6%	6482 A	0.3%
		12418 C	0.4%	3104 C	1.2%	15284 C	0.1%
				3105 T	2.1%		
				4796 T	1.0%		
				6419 C	26.1%		
				8163 T	2.1%		
				9753 T	2.4%		
				14188 T	0.4%		
				15259 T	1.9%		
				15877 T	0.5%		
F.P. = 6		F.P. = 8		F.P. = 16		F.P. = 8	



# Variant Calls – False Positives

- HV2 Homopolymer associated alignment issue
- Software has difficulty aligning around insertions and deletions

rCRS Reference Sequence



This sample has three C insertions in HV2  
309.1, 309.2, and 315.1







# Summary

- SRMs 2392 and 2392-I were successfully sequenced on four platforms
  - Sequence variants certified by Sanger were confirmed for 2392 Component B
    - G/A heteroplasmy was detected at position 1393 (minor component  $\approx 17\%$ )
  - Other SRM components are still under analysis
- $\approx 16,569$  bases measured correctly – average of 6 false positives per run
  - False positive rate = 0.04%
  - Without in-depth bioinformatics analysis
- Alignment is challenged by insertions/deletions/homopolymers
  - HV1 & HV2 C-stretch homopolymers not aligned according to forensic rules
  - Forensic mitochondrial DNA specific pipeline/solution is needed
- Software settings affect the sensitivity to minor component variants
  - Low sensitivity variant calling threshold increases false positives



# Acknowledgements

Thanks!

My colleagues at NIST

Applied Genetics Group

Pete Vallone

Multiplexed Biomolecular Science Group

Jenny McDaniel

Justin Zook



Funding from the  
**FBI Biometrics**  
**Center of**  
**Excellence 'Forensic**  
**DNA Typing as a**  
**Biometric Tool'**

# Questions?



My contact info:

email: [Kevin.Kiesler@NIST](mailto:Kevin.Kiesler@NIST)

Phone: (301) 975-4306