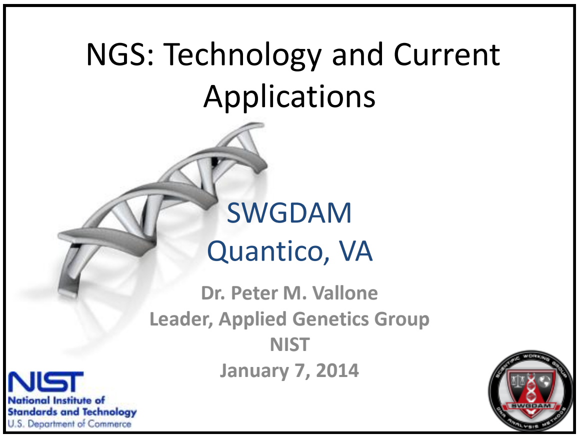
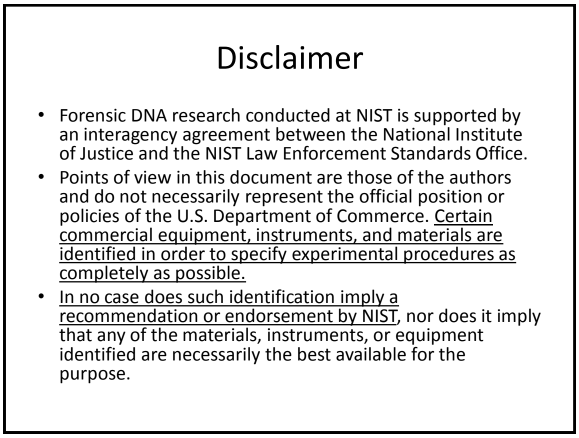


NGS: Technology and Current Applications

SWGDAM
Quantico, VA

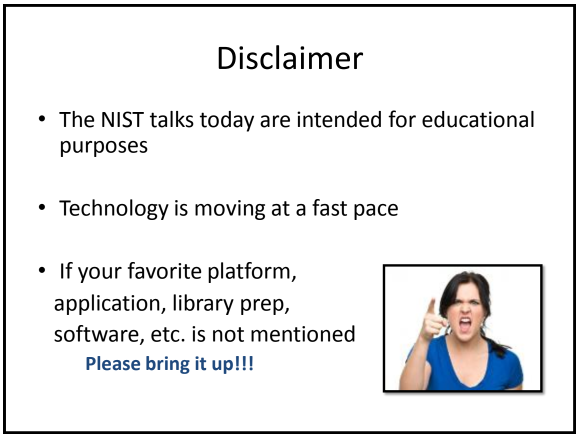
Dr. Peter M. Vallone
Leader, Applied Genetics Group
NIST
January 7, 2014

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



Disclaimer

- Forensic DNA research conducted at NIST is supported by an interagency agreement between the National Institute of Justice and the NIST Law Enforcement Standards Office.
- Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Commerce. Certain commercial equipment, instruments, and materials are identified in order to specify experimental procedures as completely as possible.
- In no case does such identification imply a recommendation or endorsement by NIST, nor does it imply that any of the materials, instruments, or equipment identified are necessarily the best available for the purpose.



Disclaimer

- The NIST talks today are intended for educational purposes
- Technology is moving at a fast pace
- If your favorite platform, application, library prep, software, etc. is not mentioned

  Please bring it up!!!

Outline

- Introduction
- Non-forensic applications
- Generalized workflow
- Platforms and throughput
- Sequencing chemistries
- Wrap up / thoughts

What's in a name???

Massively parallel sequencing

NGS

Second-generation sequencing

Next-generation sequencing

Whole-genome sequencing

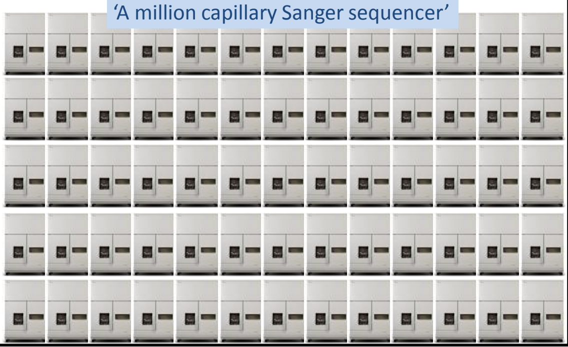
Third-generation sequencing

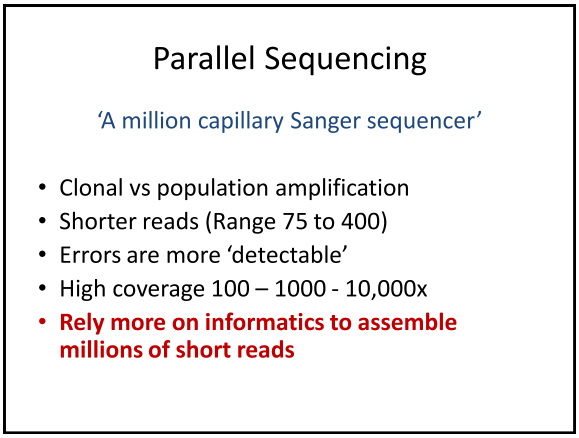
HIGH-THROUGHPUT SEQUENCING

Next-generation genomics

Parallel Sequencing

'A million capillary Sanger sequencer'

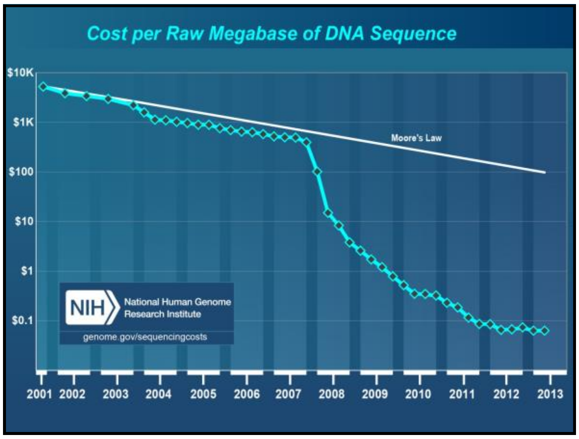




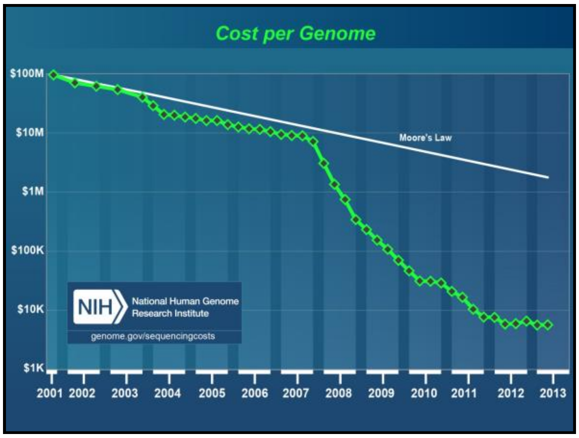
Parallel Sequencing

'A million capillary Sanger sequencer'

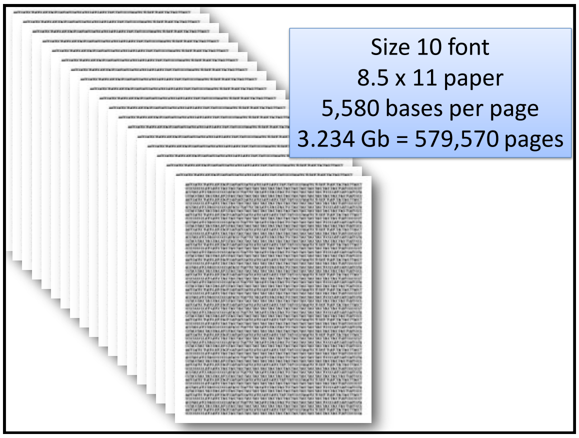
- Clonal vs population amplification
- Shorter reads (Range 75 to 400)
- Errors are more 'detectable'
- High coverage 100 – 1000 - 10,000x
- **Rely more on informatics to assemble millions of short reads**



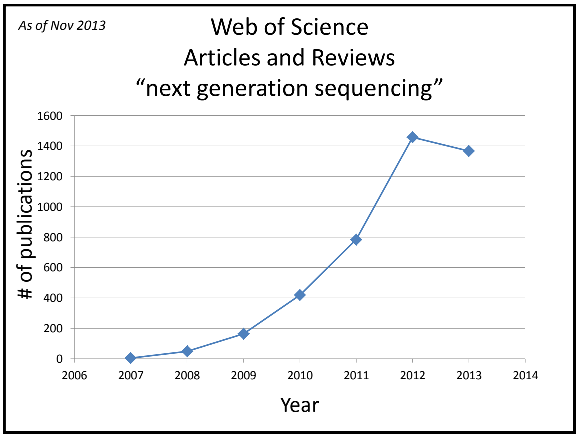
Cost per Raw Megabase of DNA Sequence



Cost per Genome

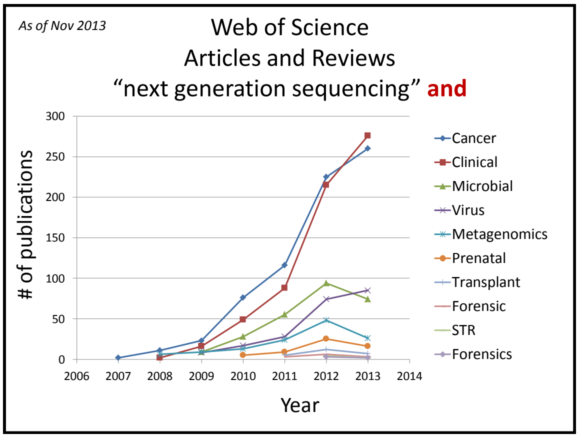


Size 10 font
8.5 x 11 paper
5,580 bases per page
3.234 Gb = 579,570 pages



As of Nov 2013

Web of Science
Articles and Reviews
“next generation sequencing”

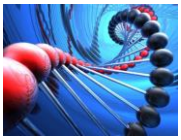


As of Nov 2013

Web of Science
Articles and Reviews
“next generation sequencing” **and**

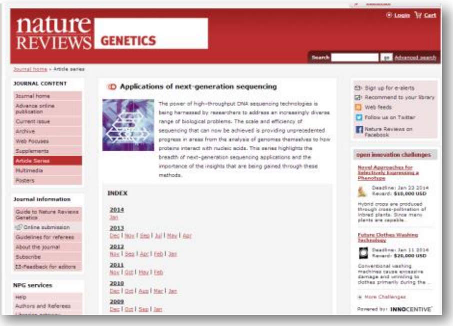
Non-forensic applications

- Clinical
- Inherited disease
- Reproductive health
- Cancer – gene fusion
- Rare variants
- Pre-implantation (genetic screening)
- Transplant medicine (HLA)
- Microbiomics/Metagenomics
- Gene expression | RNA seq
- Public health
- Ancient DNA
- NIPT (non-invasive prenatal testing)



Non-forensic applications

- Scanning Nature Reviews Genetics



nature REVIEWS GENETICS

Login Cart

Search go Advanced search

Journal home › Article series

JOURNAL CONTENT

Journal home
Advance online publication
Current issue
Archive
Web Focuses
Supplements
Article Series
Multimedia
Posters

Journal information

Guide to Nature Reviews Genetics
Online submission
Guidelines for referees
About the journal
Subscribe
Feedback for editors

NPG services

Help
Authors and Referees

Applications of next-generation sequencing

The power of high-throughput DNA sequencing technologies is being harnessed by researchers to address an increasingly diverse range of biological problems. The scale and efficiency of sequencing that can now be achieved is providing unprecedented progress in areas from the analysis of genomes themselves to how proteins interact with nucleic acids. This series highlights the breadth of next-generation sequencing applications and the importance of the insights that are being gained through these methods.

INDEX

2014
Jan

2013
Dec | Nov | Sep | Jul | May | Apr

2012
Nov | Sep | Apr | Feb | Jan

2011
Nov | Oct | May | Feb

2010
Dec | Oct | Aug | Mar | Jan

2009
Dec | Oct | Sep | Jan

Sign up for e-alerts
Recommend to your library
Web feeds
Follow us on Twitter
Nature Reviews on Facebook

open innovation challenges

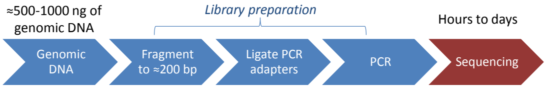
Novel Approaches for Selectively Expressing a Phenotype
Deadline: Jan 22 2014
Reward: **$10,000 USD**
Hybrid crops are produced through cross-pollination of inbred plants. Since many plants are capable...

Future Clothes Washing Technology
Deadline: Jan 11 2014
Reward: **$20,000 USD**
Conventional washing machines cause excessive damage and wrinkling to clothes primarily during the ...

More Challenges

Powered by: INNOCENTIVE

Generalized NGS Workflow



≈500-1000 ng of genomic DNA

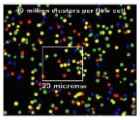
Library preparation

Hours to days

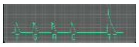
Genomic DNA → Fragment to ≈200 bp → Ligate PCR adapters → PCR → Sequencing

One template per bead/droplet/spot

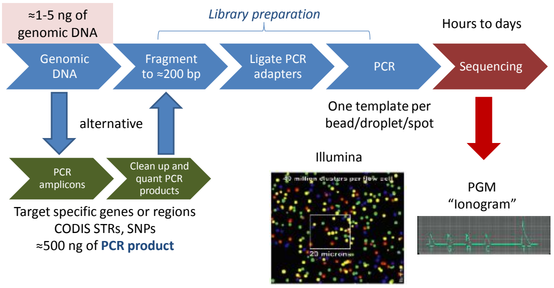
Illumina



PGM "Ionogram"



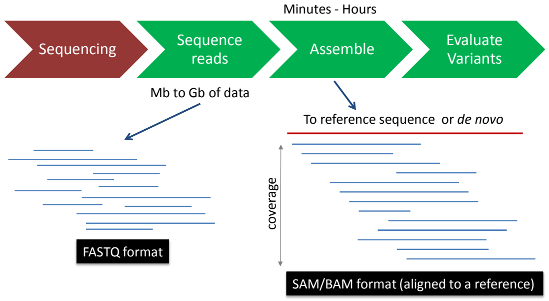
Generalized NGS Workflow



Whole Genome versus Targeted

- Whole genome
  - Genomic DNA sheared and sequenced
    - 500-1000 ng of DNA template
- Targeted
  - PCR amplified or hybridization captured regions of the genome are sheared and sequenced
    - Start with 1-5 ng of DNA -> amplify/enrich to 500-1000 ng

Generalized NGS Workflow



FASTQ Format

- FASTQ - normally uses four lines per sequence.
- Line 1 begins with a '@' character and is followed by a sequence identifier and an *optional* description
- Line 2 is the raw sequence letters.
- Line 3 begins with a '+' character and is *optionally* followed by the same sequence identifier (and any description) again.
- Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.
-
```
(1)@SEQ_ID
(2)GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCC
(3)+
(4)!''*((((***+))%%%++)(%%%%).1***-+*''))**
```

http://maq.sourceforge.net/fastq.shtml

Aligning Sequencing Reads

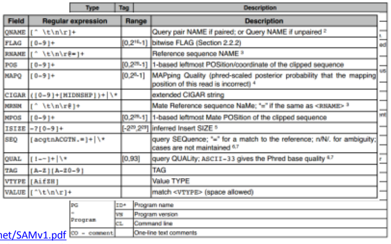
- One common algorithm is BWA
  - Burrows-Wheeler Aligner
  - Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60.
  - Li H. and Durbin R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler Transform. Bioinformatics, Epub.

Considerations when choosing an alignment software
- Speed
- Memory
- Accuracy
- Variant calling (SNPs, InDels)

SAM/BAM Format

- SAM Sequence Alignment Map Format
- Simple, tab-delimited text file
- BAM (optional compressed binary encoding)

Alignment

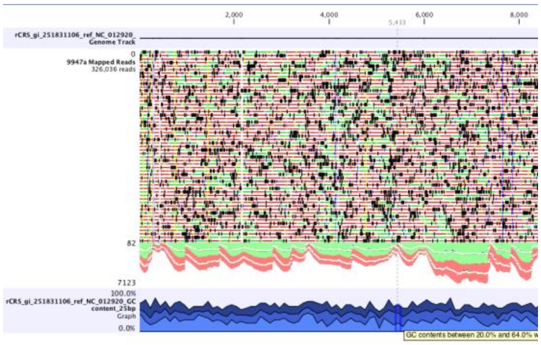


<table>
<tr><th>Field</th><th>Regular expression</th><th>Range</th><th>Description</th></tr>
<tr><td>QNAME</td><td>[^ \t\n\r]+</td><td></td><td>Query pair NAME if paired; or Query NAME if unpaired [2]</td></tr>
<tr><td>FLAG</td><td>[0-9]+</td><td>$[0,2^{16}-1]$</td><td>bitwise FLAG (Section 2.2.2)</td></tr>
<tr><td>RNAME</td><td>[^ \t\n\r@=]+</td><td></td><td>Reference sequence NAME [3]</td></tr>
<tr><td>POS</td><td>[0-9]+</td><td>$[0,2^{29}-1]$</td><td>1-based leftmost POSition/coordinate of the clipped sequence</td></tr>
<tr><td>MAPQ</td><td>[0-9]+</td><td>$[0,2^{8}-1]$</td><td>MAPping Quality (phred-scaled posterior probability that the mapping position of this read is incorrect) [4]</td></tr>
<tr><td>CIGAR</td><td>([0-9]+[MIDNSHP])+|*</td><td></td><td>extended CIGAR string</td></tr>
<tr><td>MRNM</td><td>[^ \t\n\r@]+</td><td></td><td>Mate Reference sequence NaMe; "=" if the same as <RNAME> [3]</td></tr>
<tr><td>MPOS</td><td>[0-9]+</td><td>$[0,2^{29}-1]$</td><td>1-based leftmost Mate POSition of the clipped sequence</td></tr>
<tr><td>ISIZE</td><td>-?[0-9]+</td><td>$[-2^{29},2^{29}]$</td><td>inferred Insert SIZE [5]</td></tr>
<tr><td>SEQ</td><td>[acgtnACGTN.=]+|*</td><td></td><td>query SEQuence; "=" for a match to the reference; n/N/. for ambiguity; cases are not maintained [6,7]</td></tr>
<tr><td>QUAL</td><td>[!-~]+|*</td><td>[0,93]</td><td>query QUALity; ASCII-33 gives the Phred base quality [6,7]</td></tr>
<tr><td>TAG</td><td>[A-Z][A-Z0-9]</td><td></td><td>TAG</td></tr>
<tr><td>VTYPE</td><td>[AifZH]</td><td></td><td>Value TYPE</td></tr>
<tr><td>VALUE</td><td>[^\t\n\r]+</td><td></td><td>match <VTYPE> (space allowed)</td></tr>
</table>

Type	Tag	Description
PG - Program	ID*	Program name
	VN	Program version
	CL	Command line
CO - comment		One-line text comments

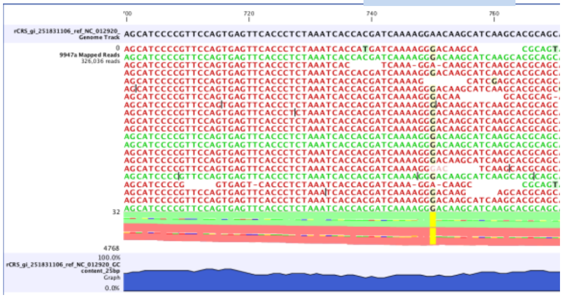
http://samtools.sourceforge.net/SAMv1.pdf
http://chagall.med.cornell.edu/NGScourse/SAM.pdf

CLC bio Sequence Viewer

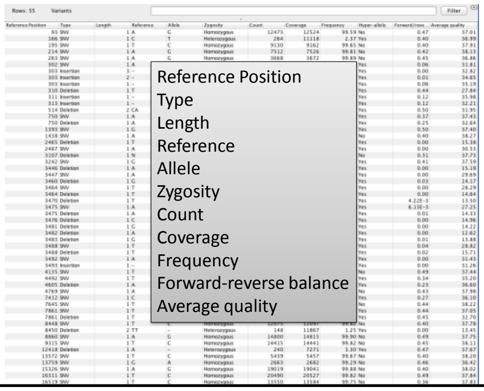

2,000
4,000
5,433
6,000
8,000
rCRS_gi_251831106_ref_NC_012920_ Genome Track
0
9947a Mapped Reads 326,036 reads
82
7123
100.0%
rCRS_gi_251831106_ref_NC_012920_GC content_25bp Graph
0.0%
GC contents between 20.0% and 64.0% w

CLC bio Sequence Viewer

Position 750 in the rCRS
A -> G Transition


700
720
740
760
rCRS_gi_251831106_ref_NC_012920_ Genome Track
AGCATCCCCGTTCCAGTGAGTTCACCCTCTAAATCACCACGATCAAAAGGAACAAGCATCAAGCACGCAGC
0
9947a Mapped Reads 326,036 reads
32
4768
100.0%
rCRS_gi_251831106_ref_NC_012920_GC content_25bp Graph
0.0%

Variant Call Table

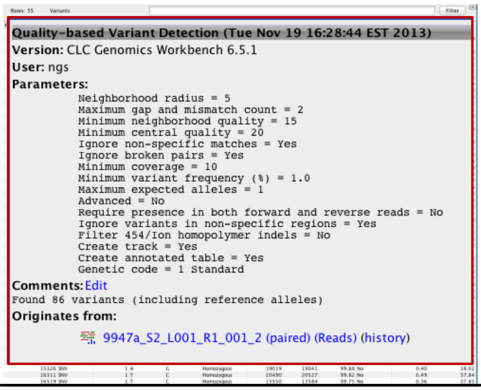


Rows: 55 Variants Filter

Reference Position	Type	Length	Reference	Allele	Zygosity	Count	Coverage	Frequency	Hyper-allelic	Forward/reve...	Average quality
93	SNV	1	A	G	Homozygous	12473	12524	99.59	No	0.47	37.01
186	SNV	1	C	T	Heterozygous	264	11118	2.37	Yes	0.40	36.99
195	SNV	1	T	C	Homozygous	9130	9162	99.65	No	0.40	37.91
214	SNV	1	A	G	Homozygous	7512	7526	99.81	No	0.42	38.13
263	SNV	1	A	G	Homozygous	3668	3672	99.89	No	0.45	36.86
302	SNV	1	A						Yes	0.06	31.81
303	Insertion	3	-						Yes	0.00	32.82
303	Insertion	2	-						Yes	0.01	34.65
303	Insertion	1	-						Yes	0.06	35.19
310	Deletion	1	T						Yes	0.44	27.84
311	Insertion	1	-						Yes	0.12	35.98
313	Insertion	1	-						Yes	0.12	32.21
514	Deletion	2	CA						Yes	0.50	31.95
750	SNV	1	A						Yes	0.37	37.43
750	Deletion	1	A						Yes	0.25	32.64
1393	SNV	1	G						Yes	0.50	37.40
1438	SNV	1	A						No	0.40	38.27
2465	Deletion	1	T						Yes	0.00	15.38
2487	SNV	1	A						Yes	0.00	30.53
3107	Deletion	1	N						No	0.31	37.73
3242	SNV	1	G						Yes	0.41	37.59
3446	Deletion	1	A						Yes	0.00	15.19
3447	SNV	1	A						Yes	0.00	29.69
3460	Deletion	1	G						Yes	0.03	14.17
3464	SNV	1	T						Yes	0.00	28.29
3464	Deletion	1	T						Yes	0.00	14.64
3470	Deletion	1	T						Yes	4.22E-3	13.50
3475	SNV	1	A						Yes	6.13E-3	27.25
3475	Deletion	1	A						Yes	0.01	14.33
3476	Deletion	1	C						Yes	0.00	14.96
3481	Deletion	1	G						Yes	0.00	14.22
3482	Deletion	1	A						Yes	0.00	12.62
3483	Deletion	1	G						Yes	0.01	13.88
3488	SNV	1	T						Yes	0.04	28.82
3488	Deletion	1	T						Yes	0.02	15.71
3492	SNV	1	A						Yes	0.00	31.43
3493	Insertion	1	-						Yes	0.00	31.26
4135	SNV	1	T						No	0.49	37.44
4492	SNV	1	T						Yes	0.34	35.20
4605	Deletion	1	A						Yes	0.23	36.80
4769	SNV	1	A						No	0.43	37.98
7412	SNV	1	C						Yes	0.27	36.10
7645	SNV	1	T						No	0.44	38.22
7861	SNV	1	T						Yes	0.44	37.05
7861	Deletion	1	T						Yes	0.45	32.70
8448	SNV	1	T	C	Homozygous	12075	12097	99.82	No	0.40	37.78
8450	Deletion	2	TT	-	Heterozygous	148	11867	1.25	Yes	0.00	13.45
8860	SNV	1	A	G	Homozygous	14800	14815	99.90	No	0.49	37.75
9315	SNV	1	T	C	Homozygous	14415	14441	99.82	No	0.45	38.11
12418	Deletion	1	A	-	Heterozygous	240	7271	3.30	Yes	0.47	37.67
13572	SNV	1	T	C	Homozygous	5439	5457	99.67	No	0.40	38.20
13759	SNV	1	G	A	Homozygous	2663	2682	99.29	No	0.46	36.42
15326	SNV	1	A	G	Homozygous	19019	19041	99.88	No	0.40	38.02
16311	SNV	1	T	C	Homozygous	20490	20527	99.82	No	0.49	37.84
16519	SNV	1	T	C	Homozygous	13550	13584	99.75	No	0.36	37.83

Reference Position
Type
Length
Reference
Allele
Zygosity
Count
Coverage
Frequency
Forward-reverse balance
Average quality

Variant Call Table



Quality-based Variant Detection (Tue Nov 19 16:28:44 EST 2013)

Version: CLC Genomics Workbench 6.5.1

User: ngs

Parameters:

```
        Neighborhood radius = 5
        Maximum gap and mismatch count = 2
        Minimum neighborhood quality = 15
        Minimum central quality = 20
        Ignore non-specific matches = Yes
        Ignore broken pairs = Yes
        Minimum coverage = 10
        Minimum variant frequency (%) = 1.0
        Maximum expected alleles = 1
        Advanced = No
        Require presence in both forward and reverse reads = No
        Ignore variants in non-specific regions = Yes
        Filter 454/Ion homopolymer indels = No
        Create track = Yes
        Create annotated table = Yes
        Genetic code = 1 Standard
```

Comments: Edit

Found 86 variants (including reference alleles)

Originates from:

9947a_S2_L001_R1_001_2 (paired) (Reads) (history)

15326	SNV	1	A	G	Homozygous	19019	19041	99.88	No	0.40	38.02
16311	SNV	1	T	C	Homozygous	20490	20527	99.82	No	0.49	37.84
16519	SNV	1	T	C	Homozygous	13550	13584	99.75	No	0.36	37.83

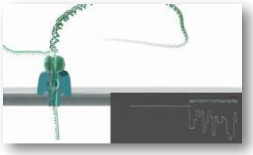
Platforms

- Illumina
  - **MiSeq**
  - HiSeq 2000/2500
  - GAIIx
- Life Technologies
  - SOLiD (5500 series)
  - **Ion Torrent PGM**
  - Ion Torrent Proton
- Pacific Biosciences
  - PACBIO RS II
- 454 Roche
  - GS jr
  - GS FLX+

October 15, 2013 – Roche shutting down 454 sequencing business
Will be phased out mid-2016

On the horizon…

- Qiagen GeneReader
  - Sequencing by synthesis approach
  - Should be available in 2014
  - QiaCube NGS (for automated library preparation)
  - Qiagen has also purchased CLC bio
- Oxford Nanopore
  - Ratcheting strand of DNA through a protein manifold
  - Bases are detected by a difference in current



Moving Targets

6 months from now these parameters will have changed

- Newer instruments
- Costs decreasing
- Throughput increasing
- Read lengths increasing
- Chemistries improving
- Library preparations – simpler/automated
- Computers faster – data storage cheaper
- Platforms leaving the market (e.g. Roche 454)
- Platforms entering the market (e.g. Qiagen GeneReader)

Low Throughput versus High Throughput

	Illumina MiSeq	Ion Torrent PGM	PacBio RS	Illumina GAIIx	Illumina HiSeq 2000
	Benchtop		High Throughput		
Instrument Cost	$128 K	$80 K	$695 K	$256 K	$654 K
Sequence yield per run	1.5-2 Gb	100-200 Mb 316 chip	100 Mb	30 Gb	600 Gb
Cost/Gb	$502	$1000	$2000	$148	$41
Run time	27 hours	2 hours	2 hours	10 days	11 days
Observed raw error rate	0.80 %	1.71 %	12.26 %	0.76 %	0.26 %
Read length	150 (300)	200 (400)	1500	150	150
Input DNA	50-1000 ng	100-1000 ng	1 ug	50-1000 ng	50-1000 ng

Adapted from: Quail et al. BMC Genomics 2012, 13:341
http://www.biomedcentral.com/1471-2164/13/341

Balancing the Equation

What question are you trying to answer?

- What instrument and/or strategy is right for my application?

- Markers
- Coverage
- Samples
- Cost (per sample and unit of information)

Other relevant questions:
- Input amounts?
- Desired level of accuracy?
- Integrity of DNA?
- Mixtures present?

Balancing the Equation

What question are you trying to answer?

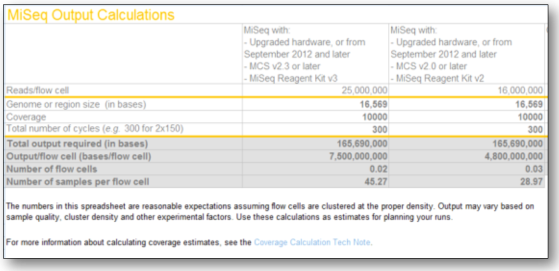
Platform 'X' provides 2 Gb of sequence per run

- Markers
  - 25 STRs and 1000 SNPs
    - 1 STR = 500 bp
    - 1 SNP = 50 bp
- Coverage
  - 600x
- Samples
  - 48

=[(25*500)+(1000*50)]*600*48
= 1.8 Gb

Balancing the Equation

What question are you trying to answer?



MiSeq Output Calculations

	MiSeq with: - Upgraded hardware, or from September 2012 and later - MCS v2.3 or later - MiSeq Reagent Kit v3	MiSeq with: - Upgraded hardware, or from September 2012 and later - MCS v2.0 or later - MiSeq Reagent Kit v2
Reads/flow cell	25,000,000	16,000,000
Genome or region size (in bases)	16,569	16,569
Coverage	10000	10000
Total number of cycles (*e.g.* 300 for 2x150)	300	300
Total output required (in bases)	**165,690,000**	**165,690,000**
Output/flow cell (bases/flow cell)	**7,500,000,000**	**4,800,000,000**
Number of flow cells	**0.02**	**0.03**
Number of samples per flow cell	**45.27**	**28.97**

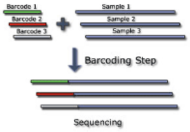
The numbers in this spreadsheet are reasonable expectations assuming flow cells are clustered at the proper density. Output may vary based on sample quality, cluster density and other experimental factors. Use these calculations as estimates for planning your runs.

For more information about calculating coverage estimates, see the Coverage Calculation Tech Note.

http://support.illumina.com/downloads/sequencing_coverage_calculator.ilmn

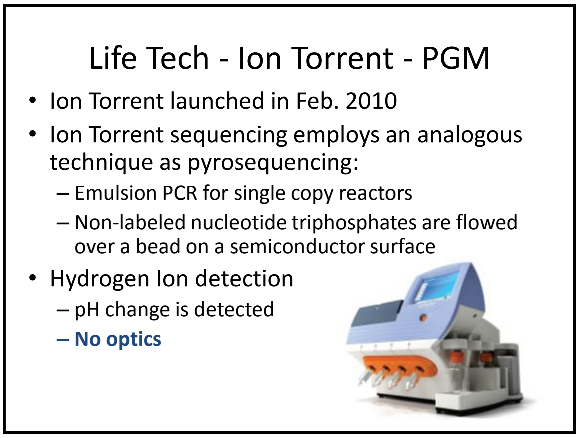
Multiplexing Samples - Barcoding

- A sample can be tagged with a unique sequence (during library preparation)
- The tagged samples could then be sequenced together and separated in the analysis stage



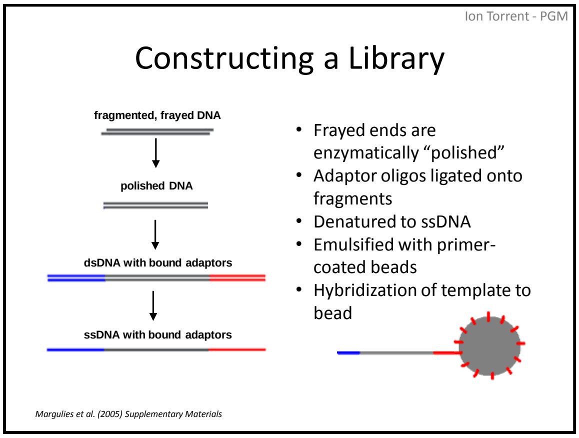
Bases in barcode index	Unique Sequence Possibilities
N	4
NN	16
NNN	64
NNNN	256
NNNNN	1024
NNNNNN	4096

Trade off volume of sequence information for more samples per run



Life Tech - Ion Torrent - PGM

- Ion Torrent launched in Feb. 2010
- Ion Torrent sequencing employs an analogous technique as pyrosequencing:
  - Emulsion PCR for single copy reactors
  - Non-labeled nucleotide triphosphates are flowed over a bead on a semiconductor surface
- Hydrogen Ion detection
  - pH change is detected
  - **No optics**

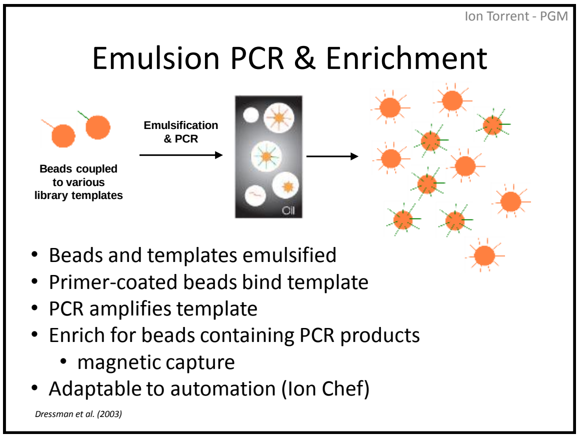


Ion Torrent - PGM

Constructing a Library

- Frayed ends are enzymatically "polished"
- Adaptor oligos ligated onto fragments
- Denatured to ssDNA
- Emulsified with primer-coated beads
- Hybridization of template to bead

Margulies et al. (2005) Supplementary Materials



Ion Torrent - PGM

Emulsion PCR & Enrichment

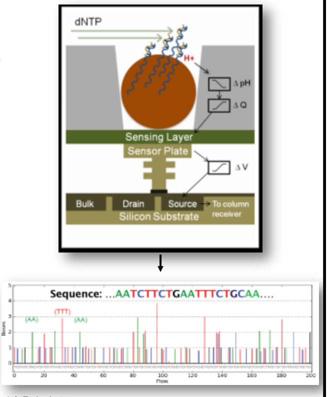
- Beads and templates emulsified
- Primer-coated beads bind template
- PCR amplifies template
- Enrich for beads containing PCR products
  - magnetic capture
- Adaptable to automation (Ion Chef)

Dressman et al. (2003)

Ion torrent PGM chip

'314' chip

- Chip flooded with one nucleotide after another
- H⁺ released when a complementary base is added to template
- Charge from the ion causes detectable pH change
- Sequencer calls the base



Life Technologies

Ion Torrent - PGM

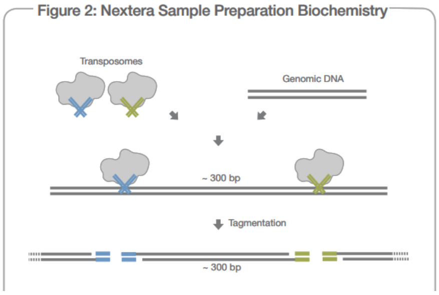
Illumina MiSeq

- MiSeq launched in Jan. 2011
- The MiSeq uses a sequencing by synthesis approach:
  - Nextera enzymatically fragments and tags DNA
  - Limited cycle PCR
  - Flow cell hybridization
  - Bridge PCR - clusters
- Fluorescent light detection
  - Each base has a unique color
  - Sequence each end of the molecule

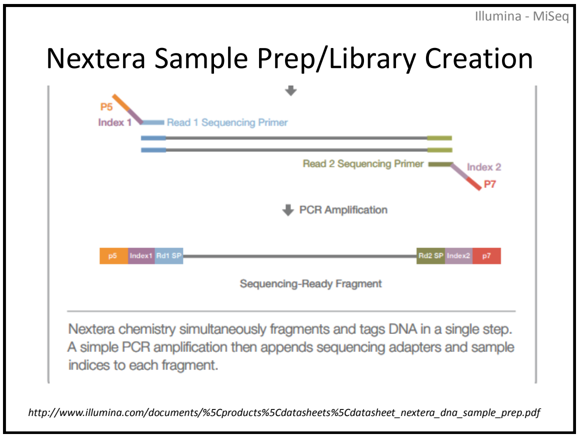
Illumina - MiSeq

Nextera Sample Prep/Library Creation

Figure 2: Nextera Sample Preparation Biochemistry



http://www.illumina.com/documents/%5Cproducts%5Cdatasheets%5Cdatasheet_nextera_dna_sample_prep.pdf

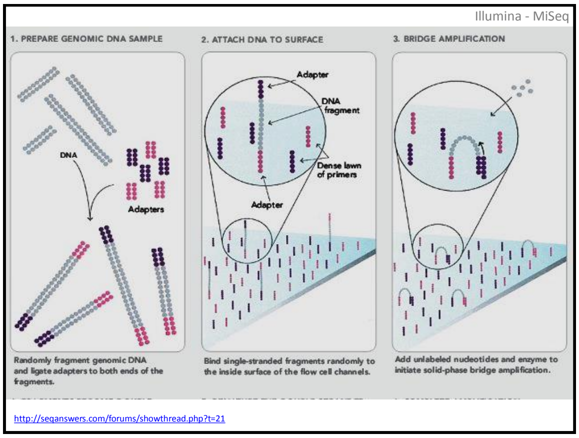


Illumina - MiSeq

Nextera Sample Prep/Library Creation

Nextera chemistry simultaneously fragments and tags DNA in a single step. A simple PCR amplification then appends sequencing adapters and sample indices to each fragment.

http://www.illumina.com/documents/%5Cproducts%5Cdatasheets%5Cdatasheet_nextera_dna_sample_prep.pdf



Illumina - MiSeq

1. PREPARE GENOMIC DNA SAMPLE

Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

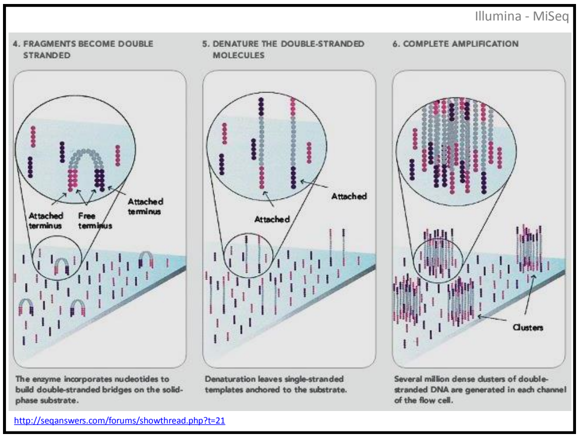
2. ATTACH DNA TO SURFACE

Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

3. BRIDGE AMPLIFICATION

Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

http://seqanswers.com/forums/showthread.php?t=21



Illumina - MiSeq

4. FRAGMENTS BECOME DOUBLE STRANDED

The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

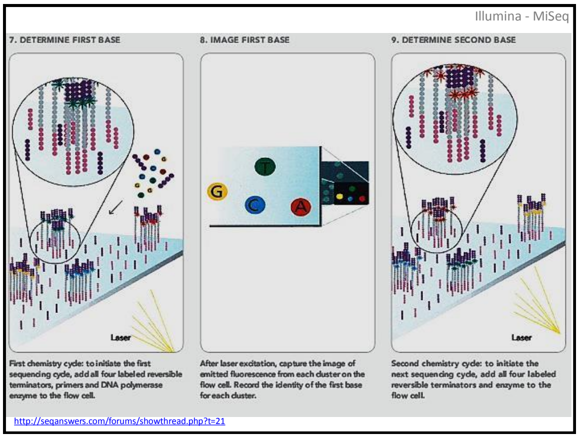
5. DENATURE THE DOUBLE-STRANDED MOLECULES

Denaturation leaves single-stranded templates anchored to the substrate.

6. COMPLETE AMPLIFICATION

Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

http://seqanswers.com/forums/showthread.php?t=21



Illumina - MiSeq

7. DETERMINE FIRST BASE

8. IMAGE FIRST BASE

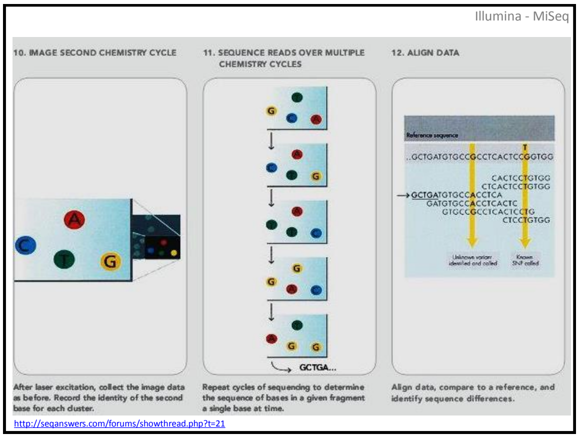
9. DETERMINE SECOND BASE

First chemistry cycle: to initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.

After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

Second chemistry cycle: to initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.

http://seqanswers.com/forums/showthread.php?t=21



Illumina - MiSeq

10. IMAGE SECOND CHEMISTRY CYCLE

11. SEQUENCE READS OVER MULTIPLE CHEMISTRY CYCLES

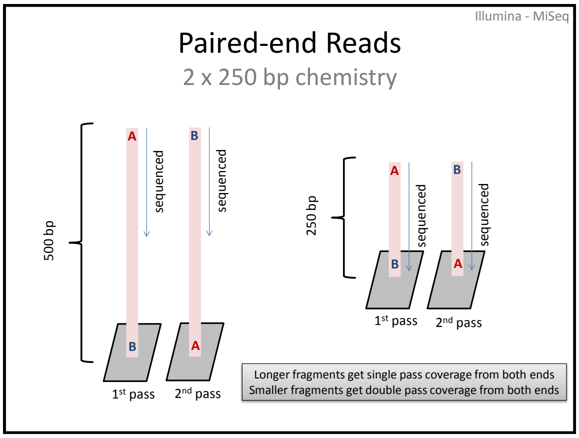
12. ALIGN DATA

After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.

Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at time.

Align data, compare to a reference, and identify sequence differences.

http://seqanswers.com/forums/showthread.php?t=21



Illumina - MiSeq

Paired-end Reads

2 x 250 bp chemistry

Longer fragments get single pass coverage from both ends
Smaller fragments get double pass coverage from both ends

Topics for further thought

- Additional genetic markers
  - SNPs (ancestry, phenotypic traits, lineage)
  - Insertion Deletion (InDels)
- Data interpretation and review – level of retention
- STR nomenclature
  - Back compatibility with existing databases
  - Future searching methods
- Ethical considerations with coding region markers
- Validation of NGS systems/methods
  - Use of existing standards (SRMs)

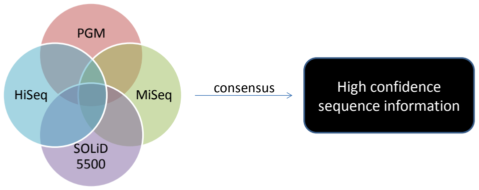
NIST SRM Support

- Further characterization of SRM 2391c, 2392, and 2392-I
- In depth sequencing of mitochondrial genomes and core STR alleles
  - Sanger
  - NGS (PGM and MiSeq)
  - Posters presented at the 25th annual ISFG meeting

"Additional Sequence Characterization of NIST SRM 2391c: PCR-Based DNA Profiling Standard"
http://www.cstl.nist.gov/strbase/pub_pres/Hill-ISFG2013-SRM2391c.pdf
"Characterization of NIST Standard Reference Materials by Next Generation Sequencing"
http://www.cstl.nist.gov/strbase/pub_pres/KieslerISFG2013poster.pdf

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



High confidence sequence information

Mitochondrial SRMs

False Positives and False Negatives

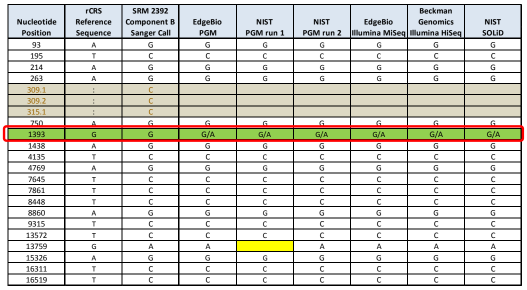
Using platform specific informatics pipeline

		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

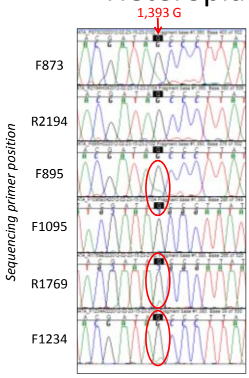
Heteroplasmy at Position 1,393

SRM 2392 Component B (9947A)

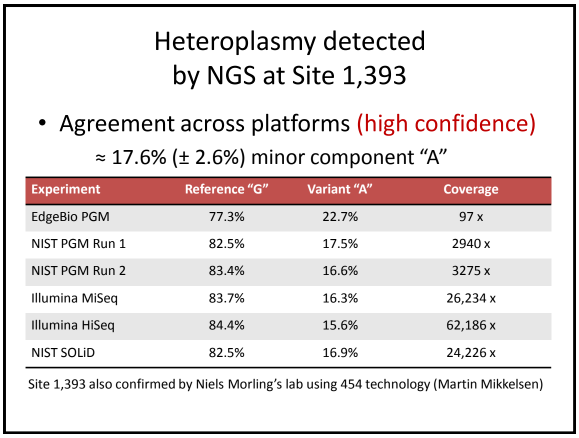


Nucleotide Position	rCRS Reference Sequence	SRM 2392 Component B Sanger Call	EdgeBio PGM	NIST PGM run 1	NIST PGM run 2	EdgeBio Illumina MiSeq	Beckman Genomics Illumina HiSeq	NIST SOLiD
93	A	G	G	G	G	G	G	G
195	T	C	C	C	C	C	C	C
214	A	G	G	G	G	G	G	G
263	A	G	G	G	G	G	G	G
309.1	:	C						
309.2	:	C						
315.1	:	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

Heteroplasmy at 1,393?



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



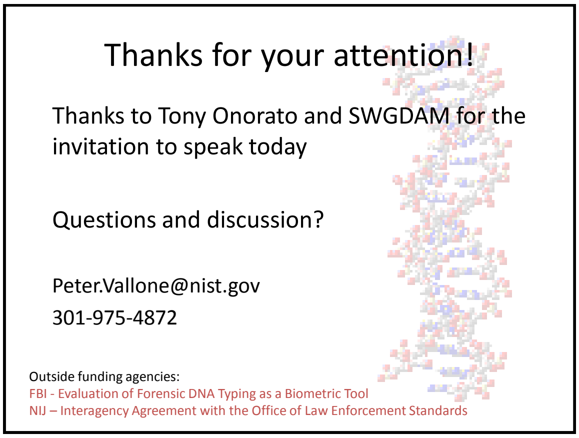
Heteroplasmy detected by NGS at Site 1,393

- Agreement across platforms (high confidence)

≈ 17.6% (± 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)



Thanks for your attention!

Thanks to Tony Onorato and SWGDAM for the invitation to speak today

Questions and discussion?

Peter.Vallone@nist.gov
301-975-4872

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