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

Next Generation Sequencing: Workflow and Infrastructure Requirements

Kevin Kiesler
Research Biologist, Applied Genetics Group

SWGDAM Meeting
January 7, 2014

Presentation Outline

- Infrastructure requirements
- Workflow
  - Front end enrichment
  - Library preparation and sequencing on PGM
  - Library preparation and sequencing on MiSeq
  - Data analysis
  - Cost analysis

What Instrument is Right for My Lab?

- What is the use case?
  - Low throughput
    - Occasional use – 1 or 2 samples per month
  - Medium throughput
    - Regular use – 10 to 20 samples per month
  - High throughput
    - Super user – 96 samples per run, infinite backlog
      - E.G. Genotyping all of China
- Keep these scenarios in mind and we will come back to this at the end for a cost analysis.

Platform Statistics

Illumina MiSeq

- **Library preparation**
  - **1 to 2 days**
- **Sequencing run time***
  - **24 to 65 hours**
- **Output per run***
  - **5 Gb to 15 Gb**
- **0.23 Gb per hour**
- **Footprint**
  - MiSeq
    - 27" x 22"

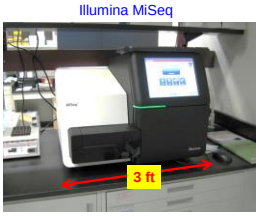
Ion Torrent PGM

- **Library preparation**
  - **2 to 3 days**
- **Sequencing run time***
  - **2.5 to 7.3 hours**
- **Output per run***
  - **100 Mb to 2 Gb**
- **0.27 Gb per hour**
- **Footprint**
  - PGM
    - 28" x 22"
  - Torrent Server
    - 9" x 20"
  - OneTouch 2
    - 12" x 16"
  - OneTouch ES
    - 11" x 14"

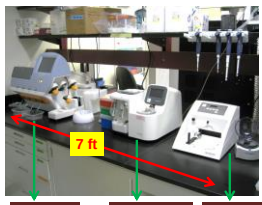
*from manufacturer's specifications

The Instruments at NIST

Illumina MiSeq



Ion Torrent PGM and accessories



PGM Sequencer | OneTouch 2 Emulsion PCR | OneTouch ES Purification

Scalable Sequencing Chemistries

MiSeq

Reagent Kit	Output
2 x 75	2.25 Gb
2 x 150	4.5 Gb
2 x 250	7.5 Gb
2 x 300	15 Gb

PGM

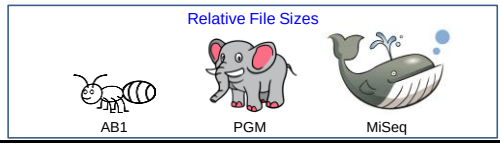
Chip	Number of Wells	Maximum Output
314	1 M	100 Mb
316	6 M	1 Gb
318	11 M	2 Gb

Additional Infrastructure

- Laboratory design
  - Downstream processes include additional PCR amplification
  - Recommendation: separate areas for
    - Pre-PCR
    - Post-PCR/Pre-library
    - Post-library amplification and Sequencing

Additional Infrastructure

- Network connectivity
  - Software updates
  - Technical support
  - Data transfer
  - Cloud computing
- Computational facilities
  - A powerful desktop system
- Data Storage
  - And lots of it!
  - For archiving
- A note on nomenclature
  - Gb = billion bases
  - GB = billion Bytes



Other Equipment You May Need

- Typical equipment
  - Pipettors
    - Single and multi-channel
  - Thermal cycler(s)
  - Vortexers
  - Shakers
  - Heat blocks
  - Magnets
  - Mini-centrifuge(s)
- Specialized equipment
  - Ultrapure water system
    - Milli-Q (Millipore)*
  - Sonicator or nebulizer
  - Size selection system
  - Agilent Bioanalyzer 2100
  - Quantitative PCR instrument
  - Qubit 2.0 Fluorometer
  - Argon supply

*May require 240 V power supply

Presentation Outline

- Infrastructure requirements
- Workflow
  - Front end enrichment
  - Library preparation and sequencing on PGM
  - Library preparation and sequencing on MiSeq
  - Data analysis
  - Cost analysis

Front End: Enrichment

- Whole genome sequencing is not yet practical
- PCR is used for targeting forensic markers
  - STRs, mtDNA, SNPs
  - We have used several approaches for mtDNA
  - We have also sequenced STRs, SNPs
- Multiplex PCR kits may become available
  - STRs, mtDNA, SNPs, Megaplex (all in one?)
- Alternative: hybridization capture
  - Sequences of interest are enriched by hybridization to DNA "baits", then pulled down by magnetic beads
  - Suitable for DNA << 200 bp in length

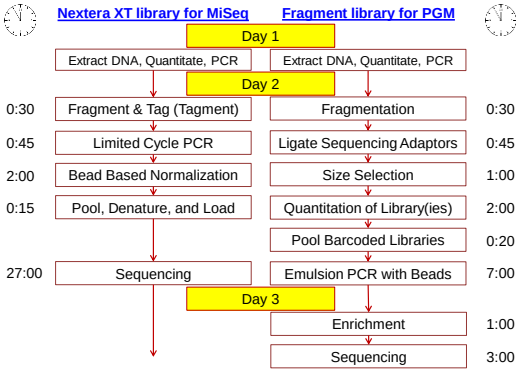
Purification and Quantitation

- Purification of PCR products
  - Removes unincorporated nucleotides and primers
  - This is necessary if using U.V. to quantify and pool PCRs
  - Removes inhibitors
- Quantitation
  - Can be done by U.V., fluorescence, or C.E.
    - Nanodrop, Qubit, or Bioanalyzer
- Pooling
  - Amplicons should be mixed in equal amounts
  - For equal sequence representation
- Multiplex PCR could eliminate these steps
  - Must be balanced

Presentation Outline

- Infrastructure requirements
- Workflow
  - Front end enrichment
  - Library preparation and sequencing on PGM
  - Library preparation and sequencing on MiSeq
  - Data analysis
  - Cost analysis

Workflow Overview



Ion Torrent PGM: Workflow Walkthrough

- The following is an example of how we have been preparing our samples for PGM sequencing
  - There are many possible protocols
  - Constantly changing & improving
  - Some alternative methods will be discussed
- The clock symbol is an estimate of how long each step will take from start to finish

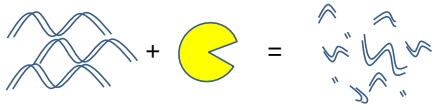
3:00

Fragmentation

PGM Step 1/8

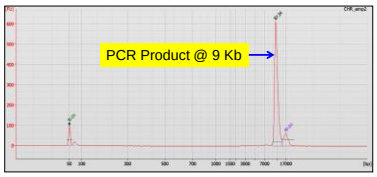
0:30

- Sequencing template must be the correct size
  - If too short, sequencing capacity is not fully utilized
  - If too long, middle of amplicon won't be sequenced
- Preferred method: enzymatic digestion
  - Special enzyme cleaves randomly

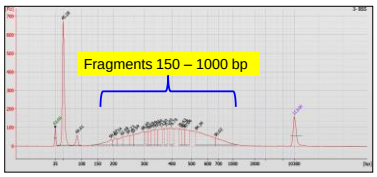


Fragmentation

Before

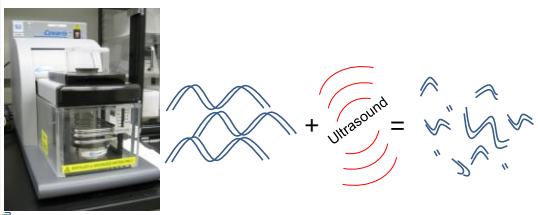


After



Fragmentation

- Alternate method: sonication
  - Not high-throughput
  - Could produce aerosolized PCR products

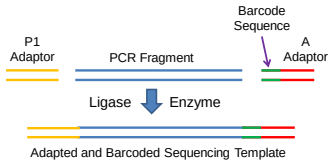


Ligation

PGM Step 2/8

0:45

- Universal sequencing adaptors are ligated to fragmented DNA
  - Adaptors termed P1 and A
- Barcode sequencing adaptors can be used in this step
  - Up to 96 samples in one run
  - 10-base sequence just downstream of the sequencing primer site



Size Selection

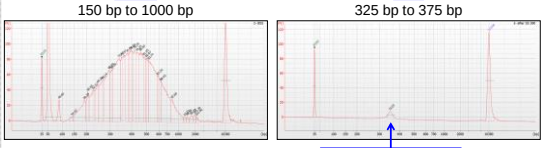
PGM Step 3/8

1:00

- Select fragments with optimal size
  - Two options
    - Invitrogen E-Gel SizeSelect System
    - Sage Science Blue Pippin
- Agilent Bioanalyzer traces below show before and after

Before
150 bp to 1000 bp

After
325 bp to 375 bp



Size selected library

Size Selection

- Invitrogen "E-Gel SizeSelect"
  - Precast agarose gels
  - Manual recovery of DNA through second tier of wells
  - Can be challenging to use during recovery step

Size Selection

- Sage Science "Blue Pippin"
  - Programmable, automatic size range collection
    - 90 bp to 50 Kb
  - Precast agarose in cassette with buffer

Quantitation of Library

PGM Step 4/8

- Next step is emulsion PCR
  - Careful quantitation is important!
  - Pooling barcoded libraries at equal amounts
  - Ratio of template to beads must be precise
    - Ideal: 10 to 30 % template positive beads
      > 30 % → polyclonal reads
- Current method: qPCR
  - Ion Library Quantitation Kit
  - TaqMan assay

2:00

- Alternative method: Bead based normalization
  - Not yet validated for fragment library prep

1:00

Emulsion PCR

PGM Step 5/8

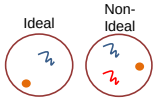
0:30

- Emulsion: nanoliter droplets of PCR reagents in oil
  - Contains DNA template and Ion Sphere Particles
  - OneTouch 2 generates emulsion and does PCR

OneTouch 2

- Ion Sphere Particles (ISP)
  - Microscopic beads with complementary sequence to adaptor P1
  - Sequencing template is attached by PCR synthesis
  - One ISP fits into one well on the Ion Chip
- Ideal: 10 % to 30 % positive beads
  - Clonal sequencing reads

Ideal Non-Ideal



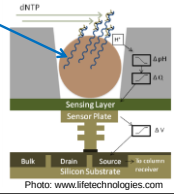


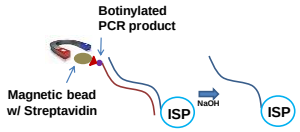
Photo: www.lifetechnologies.com

Enrich for Template Positive ISPs

PGM Step 6/8

0:30

- Automated cleanup using "ES" instrument
  - ISPs without template are removed
  - One primer in emPCR is biotinylated
  - Streptavidin beads used to bind and wash
  - Single stranded sequencing template + bead released by denaturing with NaOH

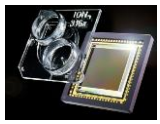


Spike In Controls and Prep PGM

PGM Step 7/8

2:30

- Control template spiked in with library
  - In case of troubleshooting
- PGM must be cleaned and initialized before run
  - This takes time
  - Proper pH is very important
- Plan run
  - Input run parameters with Torrent Server software
- Load Ion Chip
  - Pipette sample into sequencing chip
  - Deposit ISP in wells

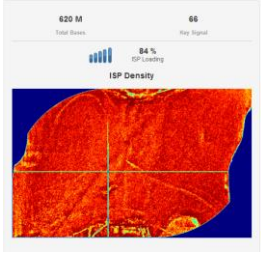


Sequencing

PGM Step 8/8

2:30

- Statistics can be monitored during run
  - Torrent Server software runs in web interface



Sequencing

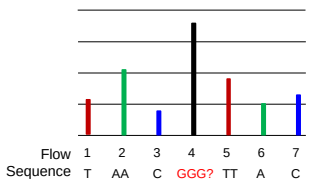
PGM Step 8/8

2:30

- Summary statistics displayed at end of run

Ion Torrent Sequencing Chemistry

- Signal is based on pH change
  - Proportional to number of bases
- Limitation: runs of 3 or more bases
  - Small amount of error adds up over several incorporations
  - Length of homopolymer stretches is difficult to measure



Presentation Outline

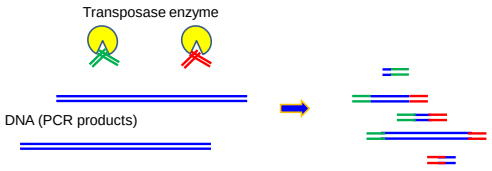
- Infrastructure requirements
- Workflow
  - Front end enrichment
  - Library preparation and sequencing on PGM
  - Library preparation and sequencing on MiSeq
  - Data analysis
  - Cost analysis

Fragmentation

MiSeq Step 1/5

0:20

- Nextera XT kit
  - Fragments and tags DNA in one step (tagmentation)

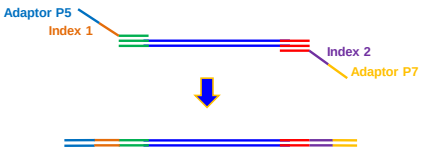


Limited Cycle PCR

MiSeq Step 2/5

0:20

- 8 cycles of PCR introduces barcodes and adaptors
  - Adaptors are complementary to flow cell oligos

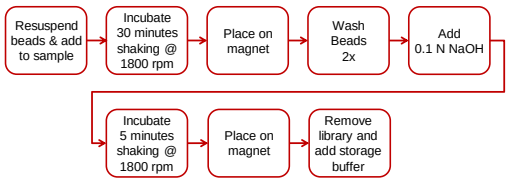


Normalize Library

MiSeq Step 3/5

1:20

- Bead based procedure normalizes quantity
  - Equal representation of pooled samples



Prepare MiSeq, Pool Libraries, Denature, and Load Cartridge

MiSeq Step 4/5

2:15

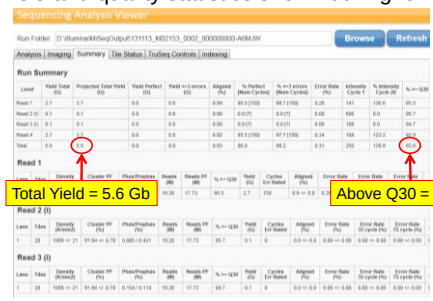
- MiSeq is cleaned before running
- Normalized libraries are pooled
  - Sequencing control (PhiX) is spiked in
- Heat denature pooled libraries (5 minutes)
- Load into cartridge in "Load Samples" reservoir

Sequencing

MiSeq Step 5/5

27:00

- Yield and quality statistics shown during run



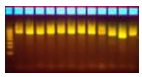
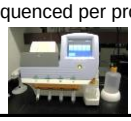
Presentation Outline

- Infrastructure requirements
- Workflow
  - Front end enrichment
  - Library preparation and sequencing on PGM
  - Library preparation and sequencing on MiSeq
  - Data analysis
  - Cost analysis

Note:
The following slides will use examples from work done at NIST with mtDNA

NIST mtDNA Sequencing

- Materials: NIST SRMs 2392 and 2392-I
  - Certified values for whole mtGenome
- PCR approach – overlapping amplicons
  - 12 PCR's (0.8 to 1.5 Kb) – on PGM
  - 3 PCR's (5 to 6 Kb) – on PGM
  - 2 PCR's (9 to 11 Kb) – on MiSeq
- Products cleaned up, quantified, and pooled
- Barcoded libraries built & sequenced per protocol
  - PGM = 316 chip
  - MiSeq = 2 x 150 chemistry


File Size and Storage

- PGM output
  - Size: about 2 GB per run
- MiSeq output
  - Size: about 11 GB per run
- Large scale storage required
  - Terabyte scale
  - Fast transfer rate
- Online "cloud" storage
  - Illumina Base Space

Example:
DROBO
Up to 48 TB

Analysis Software – Many Choices

- On platform
  - Torrent Server
  - MiSeq Reporter
- Third party software solutions
  - CLC bio Genomics Workbench
  - Softgenetics NextGENe
  - Avadis NGS (Strand Scientific Intelligence)
  - Genomatix Software Suite
  - JMP Genomics
  - Lasergene Genomics Suite (DNA Star)
- Open source software - These are command line driven programs
  - Widely used: SAMtools, Burrows-Wheeler Aligner (BWA), and Genome Analysis Tool Kit (GATK), Integrative Genomics Viewer (IGV)
  - Less commonly used: ABySS, ALLPATHS, BFAST, Bowtie, EagleView, Edena, ELAND, Euler-SR, Exonerate, Galaxy, GenomeMapper, GMAP, Gnumap, LookSeq, MAQ, MIRA2, Mosaik, MrFAST, MrsFAST, MUMmer, Novocraft, PASS, PolyBayesShort, PyroBayes, RMAP, SEQAN, SeqMap, SHARCGS, SHRiMP, Slider, SOAP, SSAHA, ssahaSNP, SSAKE, SOCS, SWIFT, SXOligoSearch, VCAKE, Velvet, Vmatch, Zoom

SOFTGENETICS
CLC bio A QIAGEN Company
DNASTAR Lasergene 11
genomatix

Items highlighted blue are in use at NIST

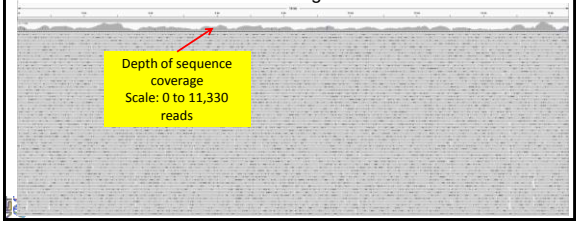
Analysis Pipeline

Sequencing → Sequence Reads (Fastq) → Assemble (BAM) → Evaluate Variants

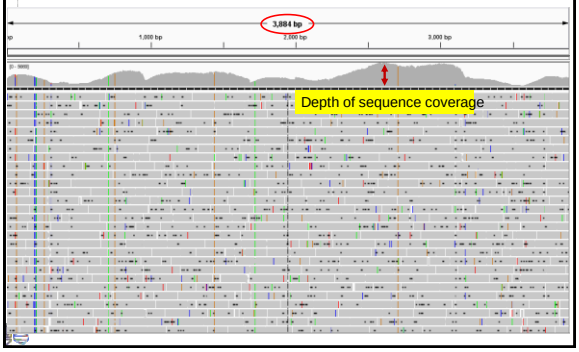
Software in use at NIST	Assembler	Variant Caller	Sequence Viewer
Platform Specific Software			
Torrent Server	X	X	
MiSeq Reporter	X	X	
Third Party Software			
NextGENe (Softgenetics)	X	X	X
Genomics Workbench (CLC bio)	X	X	X
Open Source Software			
Burrows-Wheeler Aligner (BWA)	X		
Genome Analysis Tool Kit (GATK)		X	
Integrative Genome Viewer (IGV)			X

Data Visualization for Quality Control

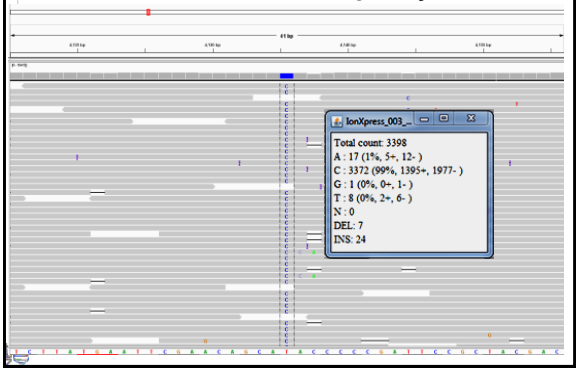
- Integrative Genomics Viewer (IGV)
  - Freely available
  - Developed by the Broad Institute
  - See: www.broadinstitute.org/IGV



Data Visualization for Quality Control

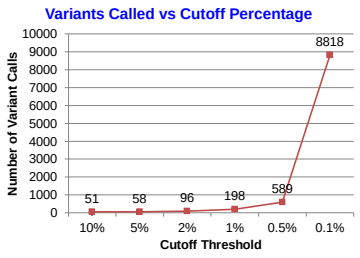


Data Visualization for Quality Control



Setting Thresholds for Variant Calls

- Increased sensitivity could result in reduced specificity
  - Number of expected mtDNA variants from Sanger sequencing = 41
  - Analysis done with NextGENe



Food for Thought

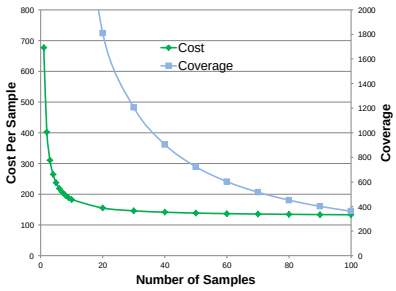
- What controls should be run?
  - Extraction reagent blanks
  - PCR no template controls
  - Acceptable threshold for background?
  - Positive control
- Phylogenetic approach for mtDNA
  - Compare variant list against known haplotypes
  - Identify errors

Presentation Outline

- Infrastructure requirements
- Workflow
  - Front end enrichment
  - Library preparation and sequencing on PGM
  - Library preparation and sequencing on MiSeq
  - Data analysis
  - Cost analysis

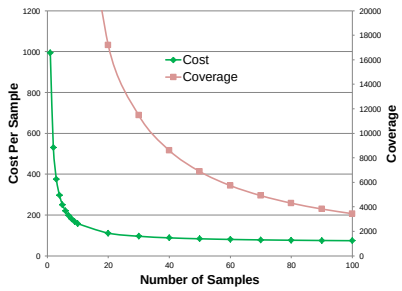
Cost and Coverage vs. Number Multiplexed

Ion Torrent PGM



Cost and Coverage vs. Number Multiplexed

Illumina MiSeq



Conclusions

- Library preparation is complex
  - Technical staff should be trained in molecular biology
  - Longer turnaround time than current methods
  - Automation will reduce complexity and time
- Cost structure best suited to higher throughput
  - Multiplexing markers and samples
- Informatics approach can have a profound effect
  - Analysis settings must be validated

Thank you for your attention!

Contact Info:
Kevin.Kiesler@nist.gov
301-975-4306