Automated NGS sample preparation and massively parallel sequencing streamlines AURORA breast cancer sample processing

Jocelyn Bush, Natalie Bir, Sean McGrath, Kristen Leraas, Anthony Bryan, Richard K. Wilson, Elaine R. Mardis, and Vincent Magrini

The Institute for Genomic Medicine

abstract

- Our automated sample preparation protocol using the Agilent Bravo liquid handling platform is able to process up to 96 samples simultaneously. Here we report on processing 193 samples for both Whole Genome Shotgun (WGS) survey sequencing and enhanced Whole Exome Sequencing (eWES) targeting deep coverage.
- The sample set was composed of normal, primary, and metastatic tumor DNA derived from blood, frozen and formalin-fixed paraffin-embedded (FFPE) tissue, respectively. The quality of the DNA ranged from high (>2 kb) to medium (0.5-2 kb) and low (<0.5 kb) based on electrophoresis gel mobility rates.
- This study highlights the processing methods for large sample WGS and eWES pooling schemes (131, 44 and 18 samples per pool) and high throughput data generation using the Illumina® NovaSeq 6000 platform.
- Unique molecular identifiers (UMIs) were incorporated into library prep to reduce read bias introduced by PCR.

background

The AURORA US: Retrospective Tissue Collection project was launched by the Translational Breast Cancer Research Consortium to focus on the evolution of breast cancer with emphasis on metastatic lesions and molecular aberrations. The WGS and eWES data we are generating and analyzing will be provided to a larger group of expert analysts as to merge with RNASeq and methylation profiling and integrate the findings with information about patients including treatment, subtype and outcomes.

Case Specimen Types:

High Quality Frozen n = 141	Medium Quality Frozen n = 1	Indeterminate Quality Frozen n = 1	Low Quality Frozen n = 0
High Quality FFPE n = 8	Medium Quality FFPE n = 40	Indeterminate Quality FFPE n = 1	Low Quality FFPE n = 1

Goal: Automated NGS library preparation for DNA of varied quality and quantity for massively parallel sequencing in a high throughput fashion

methods

Organized DNA extracts based on quantity and quality when planning high throughput processing

↓

Automated NEBNext Ultra™ II FS library prep on the Agilent Bravo liquid handing platform

↓

Quality control spot checks from each run to estimate appropriate library amplification conditions

↓

Pooled libraries based on initial DNA quality and quantity

↓

Hybridization capture with IDT xGen Lockdown® enhanced exome system

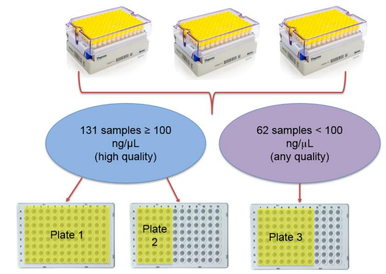
↓

Screened WGS and exome library pools on MiniSeq to determine library representation within the pool

↓

Sequenced on NovaSeq 6000

Sample Segregation Strategy for Automated Library Prep

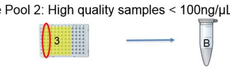


Hybridization Capture

Final Libraries Separated into 3 pools:

- Capture Pool 1: High quality samples ≥ 100ng/µL (n = 131)

- Capture Pool 2: High quality samples < 100ng/µL (n = 18)



- Capture Pool 3: Lower quality samples (n = 44)

results

Current Analysis Strategy

- Aligned reads are grouped based on UMI sequence.
- Reads in each UMI group are ranked according to per read base quality.
- The highest quality read is used as the aligned read.
- Example: 14 reads are grouped with the UMI **CGTACGTACGGT**. The read with the highest mapping score is selected as the representative read. This read lacks a variant. Is this true?

Analysis Example:

UMI Sequence	UMI Count	Highest Quality Sequence from Target
ACGTACGTACGT	20	
CGTACGTACGGT	14	
GTACGTACGTGT	16	

Coverage = total # of reads at target
★ = variant detected

Analysis Considerations

- The goal is to monitor somatic changes between primary and metastatic tissues.
- Will this strategy call attention to false positive/negative variants calls? See purple read in example.
- To minimize false calls, we are currently working on tools and conditions that will generate a representative consensus read from the grouped UMI reads.

Survey WGS Results

- Average PCR duplication rate = 3% of the total reads obtained
- With an average of 8.3X reads (± 2.74), 84% of samples achieved the target of 5X alignment coverage

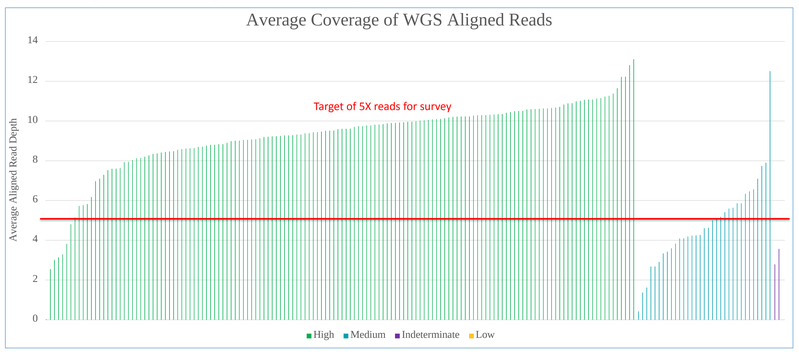


Figure 1: Sequencing of NEBNext Ultra II FS libraries processed by automated library prep was performed with one S1, one S2 and three S4 NovaSeq 6000 Reagent Kits (300 cycles). Libraries were analyzed and plotted according to initial DNA quality and their average read depth. Reads were aligned to GRCh 37 and de-duplicated using UMIs.

eWES Results

- Average PCR duplication rate = 43% of the total reads obtained
- Factors impacting coverage include quality, quantity, and complexity.

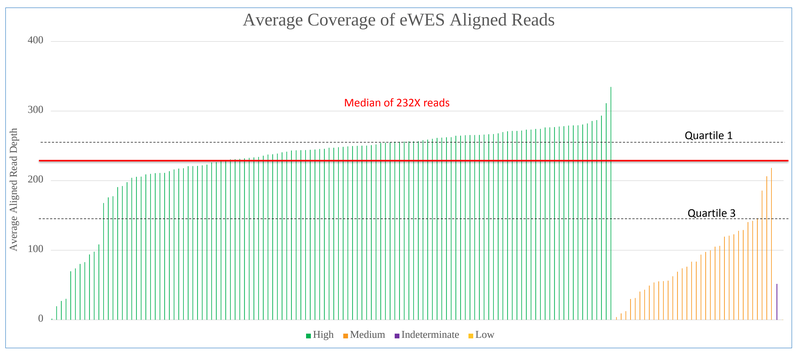


Figure 2: Sequencing of IDT xGen Hybridization and Capture libraries was performed with four S4 NovaSeq 6000 Reagent Kits (300 cycles). The libraries were analyzed and plotted according to reported DNA quality and their average read depth. Reads were aligned to GRCh 37 and de-duplicated using UMIs.

discussion

Automated library prep is suitable for all sample types, as long as quantity and quality are considered.
- Separate high quality frozen specimens from lower quality frozen and FFPE samples

QC spot checks were sufficient for estimating overall performance of a set of libraries before proceeding to hybridization capture and sequencing.

Evaluation of sequencing metrics from MiniSeq screening runs allowed for timely re-processing of capture libraries before loading on the NovaSeq.

Incorporation of UMIs during library prep allowed for PCR duplicates to be removed during sequence data alignment, thereby only retaining high quality sequences that are ready for downstream variant calling of sensitive AURORA samples.

references

- AURORA US / TBCRC Task Order, funded by Johns Hopkins University / Breast Cancer Research Foundation

acknowledgements

- Dr. Julie Gastier-Foster, Principal Investigator
- Jay Bowen, Director
- Sara Coppens, Lead Clinical Outreach Coordinator
- Kyle Voytovich and Ben Kelly, bioinformatics scientists

Contact: Jocelyn.Bush@nationwidechildrens.org