Illumina Next Generation Sequencing

**Overview**

The Sequencing and Microarray Facility (SMF) offers massively parallel next generation sequencing services on two Illumina HiSeq2000 Sequencers.  The HiSeq2000 is Illumina’s newest and most advanced sequencing platform.  It operates on Illumina’s well-established reversible terminator-based sequencing by synthesis chemistry and generates more than 500 gigabases (typically 550Gb-600Gb) of sequence per instrument run (100 nucleotides paired end).

The SMF provides comprehensive next generation sequencing services. Investigators provide the facility with genomic DNA, total RNA or ChIP DNA  (depending on the requested application) and the facility provides complete sample processing.

**The Technology and Workflow**

The HiSeq2000 workflow can be divided into four parts: **library preparation**, **cluster generation**, **sequencing by synthesis** and **data analysis**.

**Library Preparation:** A NGS library is made up of random fragments that represent the entire sample. It is created by shearing DNA (Covaris S220) into 150-400 bp fragments. These fragments are ligated to specific adapters. Library fragments of the appropriate size are then selected (size is application dependent) and isolated.  Following a sample cleanup step, the resultant library is quantified by qPCR and checked for quality using the Agilent Bioanalyzer. The SMF has automated library preparation for most applications using the Beckman SPRIworks system.

**Cluster Generation:** Library fragments are bound to a flow cell by hybridizing the fragments to a lawn of oligonucleotides complementary to the adapter sequences. Bound fragments are clonally amplified by bridge amplification to create millions of individual dense clusters of clones.  Cluster generation occurs in a closed environment on the Illumina cBOT instrument.

**HiSeq2000 Sequencing:** Sequencing on the flow cell employs Illumina’s well-established sequencing-by-synthesis chemistry. This chemistry utilizes four reversible terminator nucleotides, each possessing a different fluorescent dye and a chemically blocked hydroxyl group.  To begin sequencing, primers are hybridized to single stranded, covalently bound templates on the flow cell. Fluorescently labeled nucleotides are then flowed across the flow cell. During chain extension the fluorescent nucleotides compete for incorporation into the growing DNA chain. A single complimentary nucleotide is incorporated into each DNA molecule, terminating the chain and resulting in the simultaneous one base extension of millions of DNA clusters.  The incorporated nucleotides are excited by a laser, and emit their characteristic fluorescence.  This fluorescence is detected and recorded in an imaging step. Following base detection the fluorescent dye is cleaved and the 3’ hydroxyl block is chemically reversed, allowing chain extension to continue. This is repeated 36 to 100 times, generating a series of images.

**Data Analysis:** The raw data generated is imaged and bases are called before sequence analysis begins. Sequences generated are de-multiplexed, aligned to a reference genome and transferred to an institutional server where it is accessed by MDACC bionformaticians. Data analysis is performed in collaboration with faculty from the Department of Bioinformatics.

**Paired End Runs**

The paired end module is used to perform sequencing from both ends of the adapter-ligated fragments. Using the paired end module fragments are first sequenced from one end then essentially flipped, and sequenced from the other direction. This doubles the amount of sequence data obtained from each cluster and may provide positional information.

**Services Provided**

The Illumina Hiseq2000 sequencer is a very flexible platform, enabling a wide variety of applications that differ only in sample preparation and downstream data analysis.

**Sample Preparation Services:**1. Library preparation - includes sample QC and quantification, sample fragmentation (Covaris ultrasonicator), library QC, qPCR library quantification and cluster generation.

2. Sample Indexing: To reduce cost we are using barcodes that identify individual samples, which are then mixed together for sequencing in a single lane. We can multiplex: 4+ exomes/lane (50x - 60x coverage) 6+ ChIP-seq/lane 2 - 8 RNAseq/lane

3. Exome/custom Target Enrichment using the Nimblegen EZ-exome , Agilent Custom Target Enrichment and the Illumina Truseq/Trusight systems 4. cDNA Synthesis from total RNA .

**Supported Applications**

The SMF provides a comprehensive NGS service. Supported applications include:

**Whole genome sequencing** of Human, Mouse, Rat, Yeast, Monkey, Viral, Bacterial and other genomes. For applications in cancer research, the SMF provides sequencing of matched tumor and normal samples.

**Transcriptome Analysis**

Transcriptome analysis may be quantitative (gene expression analysis) and/or qualitative (transcript discovery, splice variant identification, coding SNP validation). The SMF offers several options for transcriptome analysis. The choice of sample preparation method is based on the investigator’s experimental objective and should be decided in conjunction with the bioinformatician.

**mRNA-Seq-** Uses oligo dT based capture for Poly enrichment followed by cDNA synthesis using random and oligo dT priming. Sequences generated map to coding regions of the genome.

**RNA-Seq-** Here rRNA depletion is performed (no Poly A enrichment) followed by cDNA synthesis utilizing oligo-d(T) and random hexamers. This method allows the sequencing of mRNA and non-polyadenylated RNA including histone mRNAs, precursors for Cajal body related small RNAs, and lncRNAs. Sequences map to exons and intergenic regions.

**Strand-specific RNA-Seq** -Preserves strand information. In addition to the information provided by traditional RNAseq, Strand-specific RNA-seq identifies antisense transcripts, determines the transcribed strand of non-coding RNAs and may help to demarcate the boundaries of overlapping genes.

**Small RNA-Seq-** Used to profile and identify changes in small RNA expression and to identify novel microRNAs.

**ChIP-Seq**- Used to identify transcription factor (protein) binding sites in genomes and specific cell types. The investigator performs chromatin IP and provides antibody captured DNA to SMF. Both ChIP Sample and mock or IgG control are required.

**Exome resequencing**: The Human Genome is comprised of approximately 3 billion base pairs,

of which only 1.2%-1.6% is coding. Exome resequencing selectively enriches for and sequences the coding regions. The SMF provides exome capture using solution based capture methods.

**Targeted Resequencing:** selectively enriches for and sequences investigator defined regions of interest. The SMF provides targeted capture using solution-based methods, long range PCR and the Nimblegen cancer panel.