**Northwest Genomics Center (NWGC)**

All sequencing is performed at the University of Washington Northwest Genomics Center (NWGC), an approved recharge center, directed by Dr. Debbie Nickerson. The NWGC has the technical staff to carry out all necessary sample processing steps for second-generation sequencing, including DNA quality control/assurance, library construction, targeted, in-solution capture methods (i.e., exome), DNA sequencer operation and maintenance, variant calling, data analysis, and IT support. The work carried out by the NWGC is done through a per assay rate cost structure that supports the operation, technical staff (i.e. non-key personnel), and reagents needed for this work.

Sample Receipt, QC and Tracking

The NWGC centralizes all receipt, tracking, and quality control/assurance of DNA samples. Samples have a detailed sample manifest (i.e., identification number/code, sex, DNA concentration, barcode, extraction method). Initial QC entails DNA quantification, sex typing, and molecular “fingerprinting” using a high frequency, cosmopolitan genotyping assay. This ‘fingerprint’ is used to identify potential sample handling errors and provides a unique genetic ID for each sample, which eliminates the possibility of sample assignment errors. Samples are failed if: (1) the total amount, concentration, or integrity of DNA is too low; (2) the fingerprint assay produces poor genotype data or (3) sex-typing is inconsistent with the sample manifest.

Library Production

Starting with minimum of 750ng of DNA, samples are sheared in a 96-well format using a Covaris LE220 focused ultrasonicator targeting 350 to 380bp inserts. This insert size improves overall library performance and allows the longer sequencing read lengths on the Illumina HiSeq X platform (150 bp) to be efficiently used without producing a significant number of overlapping reads. The resulting sheared DNA is cleaned with Agencourt AMPure XP beads to remove sample impurities prior to library construction. A two-sided AMPure cleanup is then performed in order to further restrict the fragment sizes to the desired range. End-repair, A-tailing, and ligation are performed as directed by KAPA Hyper Prep Kit without amplification (KR0961 v1.14) protocols. A final AMPure cleanup is performed after ligation in order to remove excess adapter dimers from the library. All library construction steps are automated on the Perkin Elmer Janus platform.

Clustering/Sequencing

Prior to sequencing, final library concentration is determined by triplicate qPCR using the KAPA Library Quantification Kit (KK4824), and molecular weight distributions are verified using the Agilent Bioanalyzer. Samples are sequenced on a HiSeq X using Illumina’s HiSeq X Ten Reagent Kit (v2.5). Cluster generation is performed on a cBot modified for use with the HiSeq X flow cells. These patterned flow cells allow for much higher cluster density and simplified image analysis due to fixed cluster location. Clustered flow cells are loaded onto the HiSeq X machine for sequencing.

Read Processing

Our processing pipeline consists of the following elements: (1) base calls generated in real-time on the HiSeq X instrument (RTA 2.7.6) ((2) demultiplexed, unaligned BAM files produced by Picard ExtractIlluminaBarcodes and IlluminaBasecallsToSam, and (3) BAM files aligned to a human reference (hg19hs37d5) using BWA-MEM (Burrows-Wheeler Aligner; v0.7.10) (Li and Durbin 2009). Read data from a flow-cell lane is treated independently for alignment and QC purposes in instances where the merging of data from multiple lanes is required (e.g., for sample multiplexing). All aligned read data are subject to the following steps: (1) “duplicate removal” is performed, (i.e., the removal of reads with duplicate start positions; Picard MarkDuplicates; v1.111) (2) indel realignment is performed (GATK IndelRealigner; v3.2-2) resulting in improved base placement and lower false variant calls, and (3) base qualities are recalibrated (GATK BaseRecalibrator; v3.2-2).

Data Analysis QC

Data QC includes an assessment of: (1) mean coverage; (2) fraction of genome covered greater than 10X; (3) duplicate rate; (4) mean insert size; (5) contamination ratio; (6) mean Q20 base coverage; (7) Transition/Transversion ratio (Ti/Tv); (8) fingerprint concordance > 99%; (9) sample homozygosity and heterozygosity; and (11) sample contamination validation. Genome completion is defined as having > 90% of the target at > 10X coverage and > 80% of the target at > 20X coverage. Typically this requires mean coverage of the target at 30X.