WHOLE GENOME SEQUENCING – GERMLINE

INTRO

Whole Genome Sequencing at NYGC has been developed on the HiSeq X using 2x150bp read length. We offer library preparation utilizing both PCR-free and PCR-based methods. The service is inclusive of sample QC, library prep, sequencing, and standard analysis. Delivery includes aligned .bam files as well as annotated SNV/indel and SV .vcf files (details below).

INPUT REQUIREMENTS

Upon receipt of samples, NYGC will perform QC first by measuring quantification by fluorescence using PicoGreen and second by measuring the integrity on the Fragment Analyzer. Investigator will be notified of samples that fall below the required total mass or that are degraded and not suitable for library preparation. Samples that do not meet the requirements may still be processed for sequencing but in that case NYGC takes no responsibility for sub-optimal results.

The sample submission requirements are as follows:

WGS PCR-FREE SAMPLE REQUIREMENTS

- A minimum of 2.5µg of unamplified, high molecular weight, RNase treated DNA is required
- Samples should be submitted in a total volume of 50ul to 100ul EB
- Samples should have absorbance values of OD260/280 1.7- 2.0 and OD260/230 >2.0
- Samples should be quantified by PicoGreen (or equivalent)

WGS PCR-based SAMPLE REQUIREMENTS

- A minimum of 500ng of unamplified, high molecular weight, RNase treated DNA is required
- Samples should be submitted in a total volume of 20ul to 25ul EB
- Samples should have absorbance values of OD260/280 1.7- 2.0 and OD260/230 >2.0
- Samples should be quantified by PicoGreen (or equivalent)

LIBRARY PREPARATION

DNA will be prepared using either the Illumina TruSeq Nano or PCR-free DNA sample preparation kit. The majority of the steps in this process will be carried out using the Caliper SciClone NGSx workstation, a robotics system developed and validated for automated library preparation. The library QC will include a measurement of the average size of library fragments using the FragmentAnalyzer, estimation of the total concentration of DNA by PicoGreen, and a measurement of the yield and efficiency of the adaptor ligation process with a quantitative PCR assay (KAPA) using primers specific to the adaptor sequence.

SEQUENCING

Sequencing will be performed on the HiSeq X instrument; libraries will be loaded onto the HiSeq X flowcell for clustering on the cBot using the instrument specific clustering protocol. The HiSeq X generates roughly 400M-425M passed filter 2x150bp sequencing reads per flow cell lane, after alignment and duplicate removal, this equates to roughly 30x mean genome coverage (for the gender specific ~2.85Gb mappable human genome). Each instrument processes two flow cells (16 lanes) simultaneously, and the run time is approximately 3.5 days. The NYGC currently operates 16 HiSeq X instruments.

NEW YORK GENOME CENTER

WHOLE GENOME SEQUENCING – GERMLINE

QUALITY CONTROL METRICS

For QC and finger printing purposes, all samples will be genotyped using the Illumina Human Core Exome SNP array. Concordance between genotyping calls using the SNP array and positions called from the sequencing data will be reviewed. Concordance metrics provide an additional safeguard against sample mix-up as well as an independent measure of sample contamination.

Assessment of the quality of the sequencing data will include multiple steps at different steps of the analysis pipeline. Following the completion of a sequencing run, a QC specialist will review the sequencing quality metrics including: number of pass filter reads per sample, base quality per cycle, percent base content per cycle, and the overall distribution of base quality scores. Additionally, the FASTQC tool kit has been implemented to automatically generate reports for each lane for base quality distribution, GC content distribution, and representation of particular k-mer sequences. If the raw sequencing data passes quality control threshold, it will be automatically placed into the alignment pipeline.

Post-alignment, Picard will be used to generate a sample specific metrics report. For whole genome sequencing, relevant metrics include alignment statistics, duplicate metrics, insert size, coverage statistics, and finally the X- and Y-chromosome sequence coverage is used to determine gender.

ANALYSIS

Steps in the NYGC WGS analysis pipeline include:

- Alignment of raw reads to GRCh37 using BWA-mem
- Picard MarkDuplicates
- GATK local indel realignment and base quality score recalibration
- Variant calling using GATK HaplotypeCaller
- Joint genotyping
- Annotations include variant effect predictions using SnpEFF; allele frequencies from 1000 Genomes project, NHLBI GO Exome Sequencing Project (ESP), Exome Aggregation Consortium (ExAC); dbSNP 142 rsIDs; conservation scores from PhyloP, GERP, PhastCons; damaging effect predictions from Polyphen2, SIFT; clinically relevant information from OMIM, ClinVar; regulatory potential scores from Regulome; gene ontology; pathway annotations from UniProt and ConsensusPathDB
- Structural variant calling using GenomeSTRiP

DELIVERABLES

The files delivered at the completion of a project include;

- BAM format file containing all passed filter reads and quality scores
- Recalibrated variant calls in VCF format
- Annotated variant calls in tab delimited text file format
- Raw structural variant GenomeSTRiP calls
- Annotated GenomeSTRiP results in extended BED format
- PDF summary report of SV call statistics
- 3 months of data storage, unless otherwise specified

TURNAROUND TIME

NYGC estimates turnaround time from the date samples pass QC in the NYGC laboratory. Typical turnaround times for projects of <500 samples is about 8 weeks dependent on the queue when samples arrive.