WHOLE EXOME SEQUENCING – GERMLINE

INTRO

Whole Exome Sequencing at NYGC can be performed using 2x125bp read length on HiSeq 2500 or using 2x100bp read length on Novaseq. We offer library preparation utilizing Agilent SureSelectXT V6 Target Enrichment System. The principle of Whole Exome Sequencing is to only sequence the coding regions of the genome by doing a DNA capture using a set of probes that are complementary to these regions. For human samples, the 60Mb kit is used which enriches the library with coding regions. Whole Exome Sequencing can also be performed on mouse samples. This service is inclusive of sample QC, library prep, sequencing, and standard analysis. Delivery includes aligned .bam files as well as annotated SNV/indel 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 based on customer decision. In that case NYGC takes no responsibility for sub-optimal results.

The sample submission requirements are as follows:

WES GERMLINE SAMPLE REQUIREMENTS

REGULAR INPUT SureSelect XT SAMPLE REQUIREMENTS:

- A minimum of 3.5µg of unamplified, high molecular weight, **RNase treated DNA** is required
- Samples should be submitted in a total volume of 50µl-100µl TE
- 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)
- If available, please submit an agarose gel image or Bioanalyzer results to verify DNA quality

LOW INPUT SureSelect XT 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 20µl- 25µl TE
- 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)
- If available, please submit an agarose gel image or Bioanalyzer results to verify DNA quality

LIBRARY PREPARATION

DNA will be prepared using either the Agilent SureSelectXT V6, 60mb Target Enrichment System in accordance with the manufacturer's instructions. 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 and estimation of the total concentration of DNA by PicoGreen.

SEQUENCING

Sequencing can be performed on the HiSeq 2500 or Novaseq instrument. The HiSeq 2500 generates roughly 200-250 million single end passed filter 2x125bp sequencing reads per flow cell lane. Novaseq S2 generates roughly 1.4-1.6 billion single end passed filter 2x100bp sequencing reads per flow cell lane.

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 exome sequencing, relevant metrics include alignment statistics, Hybrid Selection metrics, duplicate metrics, insert size, coverage statistics.

ANALYSIS

Steps in the NYGC WGS analysis pipeline include:

- Alignment of raw reads to GRCh37 using BWA-mem
- Picard for duplicate marking
- 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

DELIVERABLES

The files delivered at the completion of a project include:

- Expected mean target coverage as specified in the service description
- >80% of bases sequenced with a quality score above Q30
- 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
- 3 months of data storage, unless otherwise specified

TURNAROUND TIME

Turnaround time for projects with <200 samples is 8 weeks from the date samples pass QC in the NYGC laboratory. If a project is greater than 200 samples, NYGC would then deliver 100 additional samples per week. Please discuss any expedited turnaround needs with your Project Manager.