WHOLE GENOME SEQUENCING – SOMATIC

INTRO

Whole Genome Sequencing at NYGC can be performed on HiSeq X or Novaseq using 2x150bp read length. We offer library preparation utilizing both PCR-free and PCR-based methods. For PCR-free method we use the Illumina Truseq DNA Sample Prep kit and for the PCR-based method we use Illumina TruSeq Nano DNA sample prep kit. Service is inclusive of sample QC, library prep, sequencing, and standard analysis. Delivery includes aligned bam files as well as annotated SNV/indel vcf files and SV/CNV 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 failures or sub-optimal results.

The sample submission requirements are as follows:

WGS PCR-FREE SOMATIC 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 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

WGS PCR-based Nano SOMATIC 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 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 Illumina TruSeq Nano or PCR-free TruSeq 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:

- Measurement of the average size of library fragments using the FragmentAnalyzer
- Estimation of the total concentration of DNA by PicoGreen
- 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 can be performed on the HiSeq X or Novaseq intruments.

HiSeq X generates roughly 400M-425 million single end 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 11 HiSeq X Ten sequencers.
Novaseq S2 flowcell generates roughly 1.4-1.6 billion single end passed filter 2x150bp sequencing reads per flow cell lane. Run time is approximately 2.5 days. NYGC currently has 5 Novaseq sequencers.

For tumor/normal pairs we recommend an average coverage of 80x for the tumors and 40x for the normal but a dedicated and experienced project manager will assist you in defining the experimental design that better fits your project.

QUALITY CONTROL METRICS

For QC and finger printing purposes, all samples will be genotyped using the Illumina Omni 2.5M 8v1.3 array. Concordance between genotyping calls using the SNP array and positions called from the sequencing data will be reviewed. For projects with matched tumor/normal samples, the concordance within sample pairs will 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, Novosort 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 Somatic analysis pipeline include:

- Alignment of raw reads to GRCh37 using BWA-aln
- Novosort to mark duplicates
- GATK local indel realignment (joint for tumor/normal pair) and base quality score recalibration
- Somatic SNV calling - muTect, Strelka, LoFreq
- Somatic indel calling - Strelka, Scalpel, Pindel (modified to subtract normal from tumor calls)
- Somatic SNV/indel filtering with common germline variants from 1000GP, ExAC and in-house blacklist of alignment/calling artifacts
- Somatic CNV and SV calling - Crest, Delly, BreakDancer, NBIC-seq
- SNV/indel annotation via SnpEff, GATK VariantAnnotator based on ENSEMBL, 1000 Genomes, ExAC, COSMIC, Cancer Gene Census, Civic (actionability) and for coding changes, effect prediction via MutationAssessor, FATHMM_SOMATIC and CHASM
- Merging of SV calls, filtering with germline SVs, annotation with gene overlap and with sequence features around SV breakpoints

DELIVERABLES

The files delivered at the completion of a project include;

- Expected mean target coverage as specified in the service description
- >75% of bases sequenced with a quality score above Q30
- BAM format file containing all passed filter reads for tumor and matched normal sample.
- Raw variant SNV/indel calls from all variant callers in VCF format
- Union file of all SNVs and indels in VCF, MAF and tab-separated formats
- CNV/SV raw caller output in caller specific format
- Processed SV in BEDPE format files filtered to three confidence levels: merged, filtered merged, high-confidence
- PDF summary report of variant call statistics, per sample and combined for all samples in a project
- 3 months of data storage unless otherwise specified

TURNAROUND TIME

Turnaround time for projects with <200 samples is 10 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.