DNA Preparation and QC

Extraction

DNA was extracted from whole blood or flash frozen post-mortem tissue using a DNA mini kit (QIAmp #51104 and QIAmp#51404, respectively) following the manufacturer's recommendations.

DNA Quantification and Purity Measurement

DNA concentration and purity were initially measured by NanoDrop™ 2000 and verified on Qubit™ 2.0 Fluorometer.

WGS Library Preparation and Sequencing

TruSeq PCR-Free

Whole genome sequencing (WGS) libraries were prepared using the Illumina TruSeq DNA PCR- free Library Preparation Kit in accordance with the manufacturer's instructions. Briefly, 1ug of DNA was sheared using a Covaris LE220 sonicator (adaptive focused acoustics). DNA fragments underwent end-repair, bead-based size selection, adenylation, and Illumina sequencing adapter ligation. Libraries were sequenced on an Illumina HiSeq X sequencer (v2.5 chemistry) using 2 x 150bp cycles.

TruSeq Nano

Whole genome sequencing (WGS) libraries were prepared using the Illumina TruSeq Nano DNA Library Preparation Kit in accordance with manufacturer's instructions. Briefly, 100ng of DNA was sheared using the Covaris LE220 sonicator (adaptive focused acoustics). DNA fragments underwent end-repair, bead-based size selection, adenylation, and Illumina sequencing adapter ligation. Ligated DNA libraries were enriched with PCR amplification (using 8 cycles). Libraries were sequenced on an Illumina HiSeq X sequencer (v2.5 chemistry) using 2 x 150bp cycles.

NYGC Quality Control

All samples undergo rigorous quality assessment using a comprehensive set of quality measures upon completion of each step of sample processing: 1) sample receipt, 2) library preparation, 3) sequencing, 4) data analysis. Quality control measures are scrutinized by the combined efforts of our Project Management, Laboratory, Sequencing Analytics, and Bioinformatics teams. Samples that do not meet our expected quality criteria are flagged and reviewed in consultation with the investigator prior to initiation of the next step of the sample processing pipeline.

Sample QC

Automated volume check

An automated volume check on investigator samples submitted in our 2D Matrix rack tubes is performed as part of our initial QC for each sample upon arrival in the laboratory. This information is matched with the sample volume information provided by the investigator to confirm that sample integrity was not compromised during shipment.

DNA quantification

Incoming nucleic acid samples are quantified using fluorescent-based assays (PicoGreen) to accurately determine whether sufficient material is available for library preparation and sequencing.

DNA integrity

DNA sample size distributions are profiled by a Fragment Analyzer (Advanced Analytics) or BioAnalyzer (Agilent Technologies), to assess sample quality and integrity. Samples that contain degraded material and/or RNA contaminants, which could affect library preparation performance, are flagged.

Samples that do not meet our initial criteria for QC undergo further review with our Project Management team in consultation with the investigator. Investigators are provided the opportunity to re-submit new or additional material for samples that do not pass our initial QC criteria.

DNA fingerprinting and quality assessment using SNP genotyping arrays

For WGS projects, an aliquot of DNA is separately aliquoted from the submission tube upon receipt for SNP array genotyping to determine DNA integrity and identity ahead of sequencing. The genotyping results are also checked for gross sample contamination, and can reveal other forms of poor sample quality prior to sequencing.

Library QC

Library Quantification

Picogreen is used to measure the total amount of DNA in the prepared library. Quantitative PCR (qPCR) uses specific oligos complimentary to Illumina's TruSeq adapters to measure the amount of adapter-ligated DNA (ligation efficiency) that is compatible with sequencing.

Library Size distribution

Size distribution profiles of the final libraries are assessed using the Fragment analyzer/Bioanalyzer. Libraries that fall outside of the expected size range and/or contain adapter dimer contaminants are flagged.

Sequencing QC

All sequencing runs are reviewed for quality by our Sequencing Analytics team. Sequencing runs that do not pass our quality criteria for each of the metrics below are flagged and reviewed in consultation with Illumina Technical Support.

% Pass Filter (PF) clusters

Library cluster efficiency should fall within the optimal range expected for instrument and flowcell type. PF percentages outside the expected range indicate either incorrect loading concentrations or problems with a particular sequencing run.

% sample de-multiplexed

All PF reads within a single lane of a flowcell are assigned to a specific barcoded library based on the indexed read. The percentage of reads within a lane that are assigned to each sample after de-multiplexing is assessed to confirm expected sample distribution within the sample pool.

of PF reads/sample

The total number of PF reads per sample must meet the expected number of reads for a given sequencing application and analysis type, as discussed upfront with the investigator. Samples that do not meet the expected number of reads are queued for additional sequencing.

% bases >Q30

To ensure the highest quality sequencing data, FASTQ data in which at least 75% (HiSeq X) or 80% (HiSeq 2500) of bases have an Illumina Quality score >30 (a Phred like score indicating an expected 99.9% base call accuracy) are selected and used in downstream analysis.

Quality by cycle

Assessment of the quality score by cycle is used to verify that the accuracy of called bases is maintained across the entire length of the sequencing read.

GC content

GC content is reflective of both sample and library type. GC content can vary between organisms, and can be an indicator of poor sequence quality attributable to biases introduced during library preparation.

K-mer content/adapter contamination

FASTQC data is examined to identify over-represented sequences in the sequencing data, including k-mers and reads that align to the Illumina adapter sequences, both of which could indicate poor library quality and result in uneven base composition.

Data Analysis QC - WGS

Alignment metrics - PF reads, PF aligned, PF unique aligned

The number of PF reads and the fraction of these that are aligned to the reference genome (PF aligned) and map to unique sequences (PF unique aligned) are reviewed. The number of PF uniquely aligned reads has a direct impact on the mean coverage obtained for a sequenced library. Libraries that yield a lower proportion of uniquely aligned PF reads could indicate poor library or sequencing run quality.

Insert size metrics - mean, median and distribution

The mean, median, and distribution of the insert size of paired-end libraries (number of bases between the 5' and 3' adapters) are examined. Smaller insert sizes lead to overlapping reads and/or sequencing into the adapter sequences, limiting the number of usable bases for mapping and downstream analysis.

% Duplication

PCR amplification during library preparation (and flow cell clustering) can give rise to the sequencing of duplicate reads. Libraries that produce a higher than expected number of duplicate reads yield reduced coverage (DNA) or higher potential bias (RNA).

WGS coverage metrics

To ensure complete, unbiased coverage of the genome, a comprehensive panel of coverage metrics is assessed, including mean genome coverage, % of genome covered (at coverage levels 1x, 10x, 20x, etc), and coverage by chromosome. Coverage calculations are based on numbers of uniquely mapped PF reads (excluding duplicate reads). The mean coverage per sample must meet the expected target coverage for a given sequencing application and analysis type as per the project definition. Samples that do not meet the expected mean coverage are queued for additional sequencing.

Recalibrated base quality

Base Quality Score Recalibration (BQSR) is part of the Genome Analysis Toolkit (GATK) Best Practices' principles for preprocessing of aligned sequencing reads and results in adjustment of the quality scores of all reads to more accurately reflect the probability of a mismatch to the reference genome and improvements to variant calling. BQSR quality scores are graphed along to original Illumina quality scores for each sample and project to assess effective quality of sequencing data generated.

Sample contamination

Contamination checks are performed on the sequencing data to verify that the sequenced library originates from a single individual/sample. Samples in which a detectable level of contamination is identified are flagged for review. This information is used to identify sample swaps and to confirm sample identity (in combination with genotyping data).

Gender concordance

Gender is confirmed by comparing the gender information obtained from the sequencing and SNP array genotyping data, to the gender specified in the sample submission form.

Concordance to SNP array

The identity of the sequenced sample is confirmed by evaluating the concordance between sequencing and genotyping data.

Variant evaluation metrics

For SNVs and Indels, we use GATK VariantEval metrics to assess variants called per sample by: 1) the ratio of novel to known variants, 2) the TiTv ratios (for SNVs), 3) the heterozygous to homozygous ratios.

Whole Genome Sequencing Germline Analysis

Whole Genome data has been processed through the NYGC automated pipeline. Paired-end 150bp reads are aligned to the GRCh38 human reference using the Burrows-Wheeler Aligner (BWA-MEM) and processed using the GATK best-practices workflow that includes marking of duplicate reads by the use of Picard tools, local realignment around indels, and base quality score recalibration (BQSR) via Genome Analysis Toolkit (GATK).

Single Nucleotide Variant Analysis

Variant discovery is a two-step process. HaplotypeCaller is run on each sample separately in gVCF mode (GATK v3.5). This produces an intermediate file format called gVCF (genomic VCF). For projects with large number of samples, gVCFs are combined by batches into merged gVCFs. gVCFs are then run through a joint genotyping step (GATK v3.5) to produce a multi-sample VCF. Variant filtration is performed using Variant Quality Score Recalibration (VQSR) which identifies annotation profiles of variants that are likely to be real, and assigns a score (VQSLOD) to each variant. Variant effects annotation is performed using SnpEff (PMID: 22728672), bcftools (http://github.com/samtools/bcftools) and in-house software. Other functional annotations include variant frequencies in different populations from 1000 Genomes project (PMID:20981092),
Exome Aggregation Consortium – ExAC(http://biorxiv.org/content/early/2015/10/30/030338), dbSNP 147 (PMID: 11125122); cross-species conservation scores from PhyloP (PMID: 15965027), Genomic Evolutionary Rate Profiling (GERP; PMID: 21152010), PhastCons (PMID: 21278375); functional prediction scores from Polyphen2 (PMID: 20354512) and SIFT (PMID: 19561590); Clinvar (http://www.ncbi.nlm.nih.gov/clinvar/); regulatory annotations from ENCODE (PMID: 15499007) and Regulome (PMID: 22955989). Variants and annotations are exported to tabular formats for the ease of downstream analysis. Additional filtration based on functional annotation is applied to extract variants with predicted effects on protein coding.