GREGoR Methods for phs003047.v1.p1, 9/1/2023 (Release 1)

Introduction to the GREGoR Dataset

The GREGoR consortium is interested in improving diagnosis and returning results so we appreciate feedback and communication about interesting findings from the GREGoR data. Please share your thoughts via email at gregorconsortium@uw.edu.

The NHGRI GREGoR (Genomics Research to Elucidate the Genetics of Rare Disease, https://gregorconsortium.org) Consortium was established in June 2021 with the goal of developing novel tools and approaches to advance the discovery of the genetic basis of rare conditions. Participant information and numerous types of molecular data are collected and generated by the GREGoR Consortium. This data is available on the NHGRI Analysis Visualization and Informatics Lab-space (AnVIL) cloud platform via dbGaP application to phs003047.

The GREGoR Data Set conforms to the GREGoR Consortium Data Model, which is designed to support comprehensive data analysis and to expand to support the addition of new data types. The Current Data Release (R01) includes family, clinical and phenotype information as well as short-read DNA (whole exome and whole genome sequencing) and RNA sequencing data. We anticipate that future releases will include long-read DNA sequencing, methylomics, proteomics and other molecular data.

The Current Data Release (R01) contains data for 2,512 participants and 990 families across two consent groups (GRU and HMB). Family and participant information includes de-identified clinical phenotype data described using standardized ontologies. Molecular data in this release includes: 2,438 aligned DNA short read sequencing files, and 231 aligned RNA short read sequencing files.

In this release, molecular data was generated and independently processed by GREGoR Consortium Research Centers. The sample preparation and bioinformatics pipelines for each of the GREGoR Consortium Research Centers are summarized in the Methods section below . Future releases will include harmonized aligned sequencing files, single sample variant files and jointly processed variant callsets.

Methods:

CNH/Invitae:

CNH/UCI/Invitae Short Read WGS Methods Summary Input Read Quality Control

Fastqs were processed by fastqc https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ 0.12.1 with default parameters. Fastqs were concurrently processed by fastp https://github.com/OpenGene/fastp 0.23.2 with the following relevant parameter settings: -q 10 --trim_poly_g --overrepresentation_analysis --overrepresentation_sampling 100 -5 -3. Paired, passing reads are passed through fastqc as above for inspection (as post-trimmed reads) in multiqc https://multiqc.info/

Read Alignment

Trimmed reads are aligned to human reference genome GRCh38 ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA_000001405.15_GRCh38/seqs_for_alignment_pipelines.ucsc_ids/GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.gz with bwa https://bio-bwa.sourceforge.net/. Aside from standard defaults, softclipping of alts (-Y) and a fixed -K are provided to stabilize output across repeated runs. Reads trimmed for residual Illumina adapter content are used for alignment. Duplicate reads are annotated and removed with GATK https://github.com/broadinstitute/gatk 4.4.0.0 MarkDuplicates. Quality scores are recalibrated with GATK 4.4.0.0 BaseRecalibrator/ApplyBQSR.

Following BQSR, aligned reads are analyzed with GATK 4.4.0.0 CollectMultipleMetrics, CollectGcBiasMetrics, and CollectWgsMetrics; alignstats https://github.com/jfarek/alignstats; 0.10; VerifyBamID https://github.com/Griffan/VerifyBamID 2.0.1 for sample contamination; and somalier https://github.com/brentp/somalier 0.2.15 for relatedness and sexcheck.

Aligned bam files are mapped to generic sample IDs and converted to lossless crams. For recordkeeping, the site's pipeline's version and reference fasta URL used for alignment are added to the bam header as a comment (@CO) tag.

RNA-sequencing (RNAseq) library preparation and data processing:

RNA samples were isolated from whole blood collected in PAXgene tubes. RNA isolation was performed using either the Qiagen PAX Gene kit (Qiagen) or the MagMAX for Stabilized Blood Tubes RNA Isolation Kit (Invitrogen) on a KingFisher system. Total RNA was quantitated by Qubit. Ribosomal/globin-depleted, stranded libraries were prepared from 250 ng total RNA using the Watchmaker RNA Library Prep Kit with Polaris depletion (Watchmaker Genomics). Sequencing was performed on the Illumina NovaSeq platform, generating paired-end 150bp reads on S4 flowcells, using standard loading of the flow cell. Targeted coverage was 100-200 million reads per sample. We aligned RNAseq reads to the human reference genome, GRCh38, using STAR[1] (PMC3530905, v2.7.10a) and gene annotation, GENCODEv41. with the

basic two-pass protocol, allowing up to 3 mismatches and a minimum aligned length of 100bp. All other parameters were set to default. Quality control was performed using QoRTs[2] (PMC4506620) to check for any failed samples. We checked for the following:

1. Total alignment rate (unique+multi-map) >80%.
2. Correct strandedness: fr-firststrand.
3. Samples with fewer than 100 million reads were flagged as low read count.
4. RIN< 6 was flagged as poor RIN value.

Reference:

1. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan;29(1):15–21. PMCID: PMC3530905.

2. Hartley SW, Mullikin JC. QoRTs: a comprehensive toolset for quality control and data processing of RNA-Seq experiments. BMC Bioinformatics. 2015 Jul 19;16(1):224. PMCID: PMC4506620.

Broad:

Human Whole Genome Sequencing PCR-Free (v1.1-v1.3):

Preparation of libraries for cluster amplification and sequencing

An aliquot of genomic DNA (350ng in 50µL) is used as the input into DNA fragmentation (also known as shearing). Shearing is performed acoustically using a Covaris focused-ultrasonicator, targeting 385bp fragments. Following fragmentation, additional size selection is performed using a SPRI cleanup. Library preparation is performed using a commercially available kit provided by KAPA Biosystems (KAPA Hyper Prep without amplification module, product KK8505), and with palindromic forked adapters with unique 8-base index sequences embedded within the adapter (purchased from Roche). Following sample preparation, libraries are quantified using quantitative PCR (kit purchased from KAPA Biosystems), with probes specific to the ends of the adapters. This assay is automated using Agilent's Bravo liquid handling platform. Based on qPCR quantification, libraries are normalized to 2.2nM and pooled into 24-plexes.

Cluster amplification and sequencing (HiSeq X)

Sample pools are combined with HiSeqX Cluster Amp Regents EPX1, EPX2 and EPX3 into single wells on a strip tube using the Hamilton Starlet Liquid Handling system. Cluster amplification of the templates is performed according to the manufacturer's protocol (Illumina) with the Illumina cBot. Flowcells are sequenced on HiSeqX utilizing sequencing-by-synthesis kits to produce 151bp paired-end reads. Output from Illumina software is processed by the Picard data-processing pipeline to yield CRAM or BAM files containing demultiplexed, aggregated aligned reads. All sample information tracking is performed by automated LIMS messaging.

Cluster amplification and sequencing (NovaSeq 6000) Sample pools are combined with NovaSeq Cluster Amp Reagents DPX1, DPX2 and DPX3 and loaded into single lanes of a NovaSeq 6000 S4 flowcell cell using the Hamilton Starlet Liquid Handling system. Cluster amplification and sequencing occur on NovaSeq 6000 Instruments utilizing sequencing-by-synthesis kits to produce 151bp paired-end reads. Output from Illumina software is processed by the Picard data-processing pipeline to yield CRAM or BAM files containing demultiplexed, aggregated aligned reads. All sample information tracking is performed by automated LIMS messaging.

Broad Stranded RNA-Seq Method:

cDNA Library Construction

Total RNA was quantified using the Quant-iT™ RiboGreen® RNA Assay Kit and normalized to 5ng/ul. Following plating, 2 uL of ERCC controls (using a 1:1000 dilution) were spiked into each sample. An aliquot of 325 ng for each sample was transferred into library preparation which uses automated variant of the Illumina TruSeq™ Stranded mRNA Sample Preparation Kit. This method preserves strand orientation of the RNA transcript. It uses oligo dT beads to select mRNA from the total RNA sample. It is followed by heat fragmentation and cDNA synthesis from the RNA template. The resultant 400bp cDNA then goes through dual-indexed library preparation: 'A' base addition, adapter ligation using P7 adapters, and PCR enrichment using P5
adapters. After enrichment the libraries were quantified using Quant-iT PicoGreen (1:200 dilution). After normalizing samples to 5 ng/uL, the set was pooled and quantified using the KAPA Library Quantification Kit for Illumina Sequencing Platforms. The entire process is in 96-well format and all pipetting is done by either Agilent Bravo or Hamilton Starlet.

Illumina Sequencing

Pooled libraries were normalized to 2nM and denatured using 0.1 N NaOH prior to sequencing. Flowcell cluster amplification and sequencing were performed according to the manufacturer's protocols using the NovaSeq. Each run was a 151bp paired-end with an eight-base index barcode read. Data was analyzed using the Broad Picard Pipeline which includes de-multiplexing and data aggregation.

University of Washington Center for Rare Disease Research:

Relevant methods are described at https://uwmendelian.org/#/instruction

The Human Genome Sequencing Center at the Baylor College of Medicine

Exome and Genome Sequencing Methods using the Illumina Platform

Exome and genome sequencing were performed at the Human Genome Sequencing Center (HGSC) at Baylor College of Medicine through the BCM-GREGoR initiative.

Sample Quality Control

Upon receipt of samples, two independent methods were used to determine the quantity and quality of the DNA before library construction including (1) Picogreen assays and (2) E-Gels. Picogreen assays were used for DNA quantification and was based on use of Quant-iT™ PicoGreen® dsDNA reagent. Semi-quantitative and qualitative "yield gels" were used to estimate DNA sample integrity. Once DNA quality was assessed, samples moved forward for the appropriate sequencing application.

Exome Sequencing

Using 500ng of DNA, dual-indexed pre-capture libraries were prepared using KAPA reagents as described in the BCM-HGSC protocol (https://www.hgsc.bcm.edu/content/protocols-sequencing-library-construction). Libraries were pooled into 10-plex pools and capture was performed using a custom HSGC exome capture reagent (36.8Mb, custom TWIST). Paired-end sequencing was performed in a format of multiplexed pools using the Illumina NovaSeq 6000 instrument. Passing criteria for exomes included ≥90X average coverage and ≥90% of bases covered to a depth of 20X or greater.

Genome Sequencing

Using 750ng of DNA, libraries were prepared with KAPA Hyper PCR-free reagents as described in the BCM-HGSC protocol. The HGSC Illumina NovaSeq 6000 instrument fleet was employed to generate 150 bp, dual indexed and paired-end sequence reads for all samples in a format of multiplexed pools. Passing criteria included ≥30X average coverage and ≥90% of bases covered to a depth of 20X or greater.

Primary Data Processing

Post-sequencing data analysis was performed using the HGSC HgV analysis pipeline which executed base calling, mapping, merging, variant calling, post-processing, and QC metric collection for all sequencing events. Finally, annotation data was added to the VCF using a suite of annotation tools "Cassandra" that brings together frequency, function, and other relevant information using AnnoVar with UCSC and RefSeq gene models, as well as a host of other internal and external data resources.

Sequence Data Quality Control

Quality control (QC) criteria were reviewed at each procedural step in the sequencing workflows. A series of QC metrics were calculated after the mapping step. Daily quality criteria included >60% Pass Filter, >90% aligned bases, <3.0% error rate, >85% unique reads and >75% Q30 bases to achieve the application-specific metrics. Additional metrics such as library insert size (mode and mean) per sample, duplicate reads, read 1 and read 2 error rates, % pair reads and mean quality scores were also monitored. To ensure sample identity and integrity the Fluidigm SNPtrace™ method for rapidly genotyping 96 SNP sites was employed to verify gender prior to sequencing and to detect contamination using the Error Rate In Sequencing (ERIS) software developed at the HGSC. Sample concordance was measured by comparing SNP Trace genotype calls for a given sample to alignment-based genotype calls from that sample. Self-concordance was reported as a fraction of genotype matches, weighted by each SNPtrace site's MAF. The concordance report includes both self-concordance and the top six next best concordant samples. Samples whose self-concordance is less than 90% or whose self-concordance is not the highest match were further evaluated for a sample-swap.

GREGoR Stanford Site:

Short Read Genome Sequencing

The majority of GSS participants have their WGS performed at Stanford Clinical Genomics Lab (CGL). However, a small proportion of participants have WGS already sequenced from different external labs at the time of enrollment to GSS. The details about their sequencing and analysis can be found in the experiment_dna_short_read and aligned_dna_short_read data tables in AnVIL.

Peripheral blood samples were collected in EDTA vacutainer tubes using standard practices and DNA was isolated using the Maxwell RSC Buffy Coat DNA Kit (Promega Corporation, Madison, WI) according to manufacturer instructions.

Library Preparation and Short-Read Sequencing

Genome sequencing library preparation was performed using the KAPA Hyper Prep Kit (Roche Sequencing and Life Science, Indianapolis, IN) and KAPA Unique-Dual Indexed (UDI) adapters (Roche Sequencing and Life Science) according to manufacturer instructions. In brief, 500 ng of genomic DNA was fragmented to ~400-450 bp and subjected to end repair and A-tailing, followed by adapter ligation, bead-based purification, and size selection. Adaptor-ligated libraries were quantified by qPCR and pooled for sequencing, which was performed on the NovaSeq 6000 (Illumina, San Diego, CA) using either S2 or S4 flow cells and 150 bp paired-end sequencing according to manufacturer instructions. Genome sequencing depth of coverage was targeted to ≥40X, consistent with the quality metrics defined by the Stanford Medicine Clinical Genomics Laboratory (see **Table 1** below).

Bioinformatics Pipeline and Variant Calling

Sequencing data was analyzed using the Stanford Clinical Genomics Laboratory genome pipeline. Demultiplexed FASTQ files were generated from binary base call (BCL) files using bcl2fastq (Illumina), and sequencing reads were mapped to Genome Reference Consortium Human Build 38 (GRCh38). Variant calling was performed using the Illumina DRAGEN Germline Pipeline v3.10.4 (currently), which employs a Germline Small Variant Caller for SNVs/indels, and both the Structural Variant (SV) caller (Manta) and CNV Caller for structural variants. Secondary analysis data was transferred to Stanford Health Care's Google Cloud project instance for archival storage, additional custom processing, and quality control (QC) monitoring for selected sequencing, alignment, and variant metrics.

Quality Control (QC) and Quality Monitoring (QM)

The Stanford Clinical Genomics Laboratory bioinformatics quality control (QC) and quality monitoring (QM) metrics are detailed below in **Table 1**, which are monitored for overall performance and to detect any deviations from expected results over time. Any deviations from the clinical QC/QM metrics among the GREGoR Stanford Site (GSS) specimens are reviewed by a Laboratory Director and either released for downstream processing (if deviations are considered minor), or the specimen is subjected to resequencing as appropriate.

Table 1. Stanford Medicine Clinical Genomics Laboratory / GREGoR Stanford Site Short-Read Genome Sequencing Quality Metrics

	Short Read Genome QC Thresholds
Quality Control (QC) Metric	**Blood**
estimated_sample_contamination	<5%
insert_length_mean	300-600
number_of_duplicate_marked_reads_pct	<30%
pct_of_genome_with_coverage_20x_inf	>92%
properly_paired_reads_pct	>90%
q30_bases_pct	>85%
total_input_reads	>800M
average_autosomal_coverage_over_genome	≥40X
mapped_reads	>800M
mismatched_bases_r1_pct	<1.5%
mismatched_bases_r2_pct	<1.5%
reads_with_mapq_40inf_pct	>85%
variants_snps_pass	3.5-5.5M
variants_ti_to_tv_ratio_pass	1.9-2.0