GREGoR Methods for phs003047, 7/2025 (Release 3)

Introduction to the GREGoR Dataset

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 Mendelian 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 Dataset conforms to the GREGoR Consortium Data Model, which is designed to support comprehensive data analysis and to support the addition of new data types. The Current Data Release (R03) includes individual, family, clinical and phenotype information as well as molecular data including short-read DNA (whole exome and whole genome sequencing); short-read RNA sequencing; long-read DNA sequencing; methylation; ATAC-seq; and optical mapping data. We anticipate that future releases will include additional types of -omics data.

The Current Data Release (R03) consists of 8,840 participants and 3,610 families across two consent groups (GRU and HMB). Participant and family information indicate 58.9% of the dataset is composed of trios or larger families and that 47% are participants affected by a rare Mendelian condition. Molecular data in R03 includes: 9,247 aligned short-read DNA files (whole exome sequencing files = 2,284; whole genome sequencing files = 6,963), 1,043 aligned short-read RNA files, 1,772 aligned long-read DNA files and 189 aligned short-read ATAC-seq files.

In this release, the majority of 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. A subset of short-read whole genomes was harmonized by the GREGoR Data Coordinating Center (DCC). The bioinformatics pipelines used by the DCC are also summarized in the Methods section below. The harmonized dataset in this release includes aligned sequencing files (CRAMs), single sample variant files (gVCFs) and a joint callset for single nucleotide variants and indels .

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

Methods:

UCI:

UCI 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:

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 48-plexes.

Exclusion Amplification and Sequencing (NovaSeq X)

Pools are denatured with Sodium Hydroxide, diluted using an Illumina-provided Pre Load Buffer, and transferred to a uniquely-barcoded 8 lane strip tube with a Hamilton Starlet liquid handler. Strip tubes are loaded into a 300 cycle NovaSeq X 25B kit and a run is initiated with a 151 base paired end, dual-indexed read structure. A 48-plex pool is sequenced across 7 lanes of this kit in order for samples to reach the deliverable of 30X mean coverage. The output of the sequencing run, cBCL files, are demultiplexed and aligned using a DRAGEN Bio-IT server. Sample chain of custody throughout these transfers is automatically captured in our LIMS system, along with lot and expiration date information for all reagents and kits used.

Broad Germline Exome v6.0

An aliquot of genomic DNA (125ng in 50µL) is used as the input into DNA fragmentation (aka 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 with Library Amplification Primer Mix, product KK8504) and with palindromic forked adapters using unique 8-base index sequences embedded within the adapter (purchased from IDT). The libraries are then amplified by 10 cycles of PCR. Enzymatic clean-ups are performed using Beckman Coultier AMPure XP beads with elution volumes reduced to 30µL to maximize library concentration. Following library construction, library quantification is performed using the Invitrogen Quant-It broad range dsDNA quantification assay kit (Thermo Scientific Catalog: Q33130) with a 1:200 PicoGreen dilution. Following quantification, each library is normalized to a concentration of 25 ng/µL, using a 10 mM Tris HCl pH 8.0 solution. All steps performed during the library construction process and library quantification process are performed on the Agilent Bravo liquid handling system.

After library construction, hybridization and capture are performed using the relevant components of IDT's XGen hybridization and wash kit and following the manufacturer's suggested protocol, with several exceptions. A single pre-hybridization pool is created The pre-hybridization pool comprised of 96 unique libraries is created by equivolume pooling of the normalized libraries as well as 5 uL of Human Cot-1 and 2 ul of IDT XGen blocking oligos. The pre-hybridization pool undergoes lyophilization using the Biotage SPE-DRY. Post lyophilization, 4 uL of custom exome bait (TWIST Biosciences) along with 13 uL of hybridization mastermix is added to the lyophilized pool prior to resuspension. An initial incubation is performed at 95ºC for 30 seconds, after which time the incubation temperature is lowered to 65ºC at which it remains overnight. Library normalization and hybridization setup are performed on a Hamilton Starlet liquid handling platform, while target capture is performed on the Agilent Bravo automated platform.

After post-capture enrichment, library pools are quantified using qPCR (automated assay on the Agilent Bravo), using a kit purchased from KAPA Biosystems with probes specific to the ends of the adapters. Based on qPCR quantification, pools are normalized using a Hamilton Starlet and sequenced on one lane of a NovaSeq S4 flow cell using the NovaSeq XP workflow with a read length of 2x151bp.

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 DRAGEN Pipeline which includes de-multiplexing and data aggregation.

Mitochondrial DNA (mtDNA) variant calling

mtDNA single nucleotide and small indel variants were called from genome sequencing data (CRAM files) using the gnomAD-mitochondria pipeline (PMID: 35074858) and from exome sequencing data (CRAM files) using the the MToolBox pipeline (PMID: 25028726). All pipelines were run as workflows (WDL) in Terra. For both data types, individual-level mtDNA VCF files and coverage files were combined and annotated using the gnomAD-mitochondria Hail pipeline (https://github.com/broadinstitute/gnomad-mitochondria) to generate mtDNA call sets for analysis. Downstream analysis methods are reported in detail in Stenton et al., *medRxiv* 2024 (https://www.medrxiv.org/content/10.1101/2024.12.22.24319370v1).

Oxford Nanopore Technology (ONT) long-read genome sequencing

One EDTA tube was supplied per individual. HMW DNA was extracted using Circulomics Nanobind CBB Big DNA Kit (NB-900-001-01) or NEB Monarch HMW DNA extraction kit for cells and blood (NEB T3050). Approximately 5 µg of isolated DNA was sheared using Diagenode Megaruptor 3, DNA fluid+ kit (E07020001). The size of sheared DNA fragments was analyzed on the Agilent Femto Pulse System using genomic DNA 165 kb kit (FP-1002-0275). Fragment size distribution of post-sheared DNA had peaked at approximately 50 kb. Small DNA fragments were removed from the sample using PacBio SRE (Short Read Eliminator) kit (SKU 102-208-300). Library preparation was carried out using ONT ligation sequencing kit V14 (SQK-LSK114). Sequencing was performed on the PromethION 48 sequencer using R10.4.1 flow cells. Each sample was used to prepare four libraries per flow cell. Flow cells were washed using the ONT wash kit (EXP-WSH004) and reloaded with a fresh library every 24 h for a total sequencing runtime of 96 h. The Napu end-to-end pipeline was run to generate diploid *de novo* phased assemblies, harmonized variant calls against the GRCh38 reference genome (merging reference-based small variant calls and assembly-based SV calls), and haplotype-specific methylation calls (see Negi et al., 2025 for further details: https://pubmed.ncbi.nlm.nih.gov/39862869/)

PacBio long-read genome sequencing

Specific methods for library preparation, sequencing, and data processing are summarized below.

Flowcell-level processing performs preliminary processing of long-read data consisting of CCS error correction and sequencing metrics (coverage, read lengths, sequence identity, etc.).

PacBio circular consensus ("CCS" aka "HiFi") sequencing

For CCS library preparation, at least 2 µg of high molecular weight genomic DNA was depleted of fragments <10kb using the PacBio SRE HT kit (103-124-500). The resulting gDNA was then sheared to ~15kB on the Hamilton Starlet system using automated pipette shearing. This was followed by SMRTbell library preparation using the PacBio HiFi prep kit 96 (103-122-600); consisting of a post-shear cleanup, DNA end repair, ligation of PacBio adapters, and nuclease treatment. Libraries were then hybridized with PacBio standard sequencing primer and bound with SPRQ sequencing polymerase using the Revio SPRQ polymerase kit 96 (103-497-000). Polymerase bound SMRTbell complexes were then normalized to 300pM and pooled to achieve their specific target coverages. CCS sequencing was performed on the Revio instrument using 25M SMRT Cells (102-202-200) and Revio SPRQ Sequencing Plate (103-504-900), with a 2-hour pre-extension time and 30-hour movie time per SMRT cell. Primary data review including quality filtering, basecalling, and demultiplexing are performed via onboard processors.

Error correction for reads generated in CCS mode was performed on-board with the vendor's ccs software[1] and settings --all --subread-fallback --num-threads 232 --streamed <movie_name>.consensusreadset.xml --bam <movie_name>.reads.bam. The resulting reads, including those that failed error correction and packaged into a single BAM[2] file for further analysis. Should a run contain individually barcoded libraries, on board demultiplexing will produce a single unique BAM file for each index present. Methylation signals called via PacBio's Jasmine software are also encoded within the BAM as tags (MM and ML).

Alignment and metrics calculation

Reads were aligned to the GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.gz[4] (herein referred to as "grch38_noalt") reference sequence using pbmm2[5] CCS: `pbmm2 align <input.bam> <ref.fa> <output.bam> --preset CCS --sample <sample_name> --strip --sort --unmapped`). The MD tag encoding mismatched and deleted reference bases were computed post-alignment using the `samtools calmd` command[6]. Genome-wide and per-chromosome coverages were computed using MosDepth[7], and read length/sequence identity metrics using NanoPlot[8] and custom scripts.

Sample-level processing combines (if necessary) aligned data for samples spanning multiple flowcells and performs variant calling (SNV, small indel, and structural variants) on each combined per-sample dataset.

Small variants (SNVs, indels < 50 bp)

Small variants (defined as SNVs and insertions and deletions less than 50 bp in length) are discovered using PEPPER-Margin-DeepVariant, which generates both a VCF and gVCF file (the latter intended for later joint variant calling in cohort mode, if necessary). For computational efficiency, callers are parallelized over each chromosome (excluding random, unplaced, decoy, alternative loci, HLA, and EBV contigs) in the reference sequence. Per-chromosome calls are merged upon completion.

Structural variants (SVs)

Structural variants are called with two reference-based algorithms, PBSV and Sniffles. Indel/structural variants were called with PBSV using the “-tandem-repeats” argument for increased sensitivity to repeat expansions/contraction variants according to the software’s documentation[10]. In accordance with guidance from the Sniffles[11] authors, indel/structural variants were called with minimum read support of 2, minimum read length of 1000 bp, and minimum mapping quality of 20. Each algorithm is run per-chromosome for computational efficiency. Per-chromosome and per-algorithm VCFs are then merged to simplify downstream analysis.

Cohort-level processing leverages data across samples to improve the discovery of lower frequency variants or to establish a multi-sample baseline against which other samples can be compared for deviation and discovery.

Small variants

Cohort-level processing for small variants is carried out with the gVCFs (per sample) output in the single-sample level, using GLnexus [13]. A single, multi-sample (g)VCF is output.

Structural variants

Cohort-level processing for structural variants is under development.

Copy number variants (CNVs)

Copy number variant detection must examine multiple samples simultaneously in order to establish a normalized coverage baseline from which to detect significant coverage deviation in samples. Thus, CNV detection is enabled only at the cohort level using the gCNV tool. Normalized coverage is computed in windowed, non-overlapping intervals across the genome. Currently, our window size is set to 100,000 bp (~10x the typical long read length). This can facilitate detection of large rearrangements, but may miss smaller copy number changes (for which we rely on per-sample structural variant detection to mark as indels).

References

1. https://github.com/PacificBiosciences/pbccs
2. Li, H. et al. "The Sequence Alignment/Map format and SAMtools" Bioinformatics. 2009 Aug 15;25(16):2078-9. doi: 10.1093/bioinformatics/btp352.
3. Pacific Biosciences, Inc. "What is in the reads.bam?" CCS Docs https://ccs.how/faq/reads-bam.html.
4. Li, H. "Which human reference genome to use?" https://lh3.github.io/2017/11/13/which-human-reference-genome-to-use
5. https://github.com/PacificBiosciences/pbmm2
6. Danecek et al. "Twelve years of SAMtools and BCFtools" Gigascience. 2021 Feb 16;10(2):giab008. doi: 10.1093/gigascience/giab008.
7. Brent S Pedersen and Aaron R Quinlan "Mosdepth: quick coverage calculation for genomes and exomes" Bioinformatics. 2018 Mar 1;34(5):867-868. doi: 10.1093/bioinformatics/btx699.
8. Wouter De Coster et al. "NanoPack: visualizing and processing long-read sequencing data" Bioinformatics. 2018 Aug 1;34(15):2666-2669. doi: 10.1093/bioinformatics/bty149.
9. Javed et al. "Detecting sample swaps in diverse NGS data types using linkage disequilibrium" Nat Commun. 2020 Jul 29;11(1):3697. doi: 10.1038/s41467-020-17453-5.
10. Aaron M Wenger et al. "Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome" Nat Biotechnol. 2019 Oct;37(10):1155-1162. doi: 10.1038/s41587-019-0217-9.
11. Fritz J Sedlazeck et al. "Accurate detection of complex structural variations using single-molecule sequencing" Nat Methods. 2018 Jun;15(6):461-468. doi: 10.1038/s41592-018-0001-7.
12. Babadi et al., "GATK gCNV: accurate germline copy-number variant discovery from sequencing read-depth data", https://github.com/broadinstitute/gatk/blob/master/docs/CNV/germline-cnv-caller-model.pdf
13. GLnexus, https://github.com/dnanexus-rnd/GLnexus

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

Long Read Whole Genome Sequencing

Genomic DNA (gDNA) was isolated from whole blood aliquots using the salting/alcohol precipitation procedure (Puregene Blood Core Kit, Qiagen). The gDNA was quantity and quality checked by Qubit and Femtopulse. Small fragments were removed using the short read eliminator kit (SRE, PacBio) and sheared using the Megaruptor 3 (Diagenode) to a desired size of 15-18kb. WGS libraries were prepared and indexed using the Smartbell prep Kit 3 (PacBio), quantitated by Qubit and a final size determination done by femtopulse.

During the whole genome library preparation, slight modifications were made to the PacBio protocol and shear settings in order to better target sequencing parameters. With these modifications sequencing depth and coverage increased to 30x and above, reaching our coverage goals and data yield preferences. The most impactful modification made was an increase in volume from the baseline of 130ul detailed in the PacBIo protocol to a volume of 150ul , which allowed gDNA samples to go into solution more freely, streamlining the size selection process. Once WGS libraries were completed, they were then polymerase bound

through the Anealing, Binding and Clean up step (ABC) detailed through SMRTlink on the revio software system. Samples undergo a 30 hour sequencing, yielding an array of HiFi data.

Sequencing data was analyzed using PacBio's HiFi WDL pipeline. The pipeline takes as input, unmapped bams generated from the Revio. The data is aligned against GRCh38, followed by SNV and SV calling by DeepVariant and PBSV respectively. This information is used with HiPhase to phase the aligned bams. For quality control, sequencing runs must meet a minimum mean depth of 25x. For quality monitoring, we record yield and mean lengths for each run to track trends over time.

GREGoR Data Coordinating Center

Harmonization of GREGoR whole genome sequencing data

Harmonization of short-read, whole genome sequencing (srWGS) data was performed by the GREGoR Data Coordinating Center using the Whole Genome Germline Single Sample WARP pipeline in DRAGEN-GATK mode (v3.1.6). This pipeline includes alignment to the GRCh38 genome reference using the DRAGMAP aligner and duplicate marking with Picard v2.26.10. Single sample variant calling was performed with GATK HaplotypeCaller using the Dragstr model and hard filtering parameters.

Joint variant calling for GREGoR srWGS was performed using the Genomic Variant Store (GVS). This pipeline was developed to perform joint calling at scale and is based on a schema designed for querying and rendering variants in which the variants are stored in GVS and rendered to an analyzable variant file format. Functional annotation of the GREGoR joint callset was performed with Variant Effect Predictor v112.