Published on *NHLBI Trans-Omics for Precision Medicine* (https://topmed.nhlbi.nih.gov)

TOPMed Whole Genome Sequencing Project - Freeze 5b, Phases 1 and 2

Updated 10/28/2021

Table of Contents

Introduction

Overview

Trans-Omics for Precision Medicine (TOPMed), sponsored by the National Heart, Lung and Blood Institute (NHLBI), generates scientific resources to enhance our understanding of fundamental biological processes that underlie heart, lung, blood and sleep disorders (HLBS). It is part of the broader Precision Medicine Initiative, which aims to provide disease treatments that are tailored to an individual's unique genes and environment. TOPMed contributes to this initiative by integrating whole-genome sequencing (WGS) and other -omics data (e.g., metabolic profiles, protein and RNA expression patterns) with molecular, behavioral, imaging, environmental, and clinical data. In doing so, the TOPMed program seeks to uncover factors that increase or decrease the risk of disease, identify subtypes of disease, and develop more targeted and personalized treatments.

Currently, TOPMed includes >70 different studies with ~145,000 samples with whole genome sequencing (WGS) completed or in progress. These studies encompass several experimental designs (e.g. cohort, case-control, family) and many different clinical trait areas (e.g. asthma, COPD, atrial fibrillation, atherosclerosis, sleep). See study descriptions under the "Studies" tab on the TOPMed web site (topmed.nhlbi.nih.gov [19]).

TOPMed WGS genotype call sets (called "Freezes") are being released on dbGaP periodically (~6-12 month intervals). WGS data for samples from Phase 1 studies, with reads mapped to human genome build GRCh37, were released in 2016 (Freeze 3) and 2017 (Freeze 4). The Freeze 5b genotype call set, with samples from Phase 1 and 2 studies and reads mapped to genome build GRCh38, are being released in 2018. A summary of the dbGaP accessions for these studies, including their approximate sample numbers, are provided in Table 1.

Some TOPMed studies have previously released genotypic and phenotypic data on dbGaP in "parent" accessions (see Table 1). For those studies, the TOPMed WGS accession contains only WGS-derived data and, therefore, genotype-phenotype analysis requires data from both the parent and TOPMed WGS accessions. For the studies in the Table without a specific parent accession number, the TOPMed WGS accession contains both genotype and phenotype data.

Table 1 : Summary of TOPMed Study Accessions in Freeze 5b

Project[1]	Study Accession	Study Name[2]	Study/Cohort Abbreviation	Study PI	Sample Size[3]	Sequencing Center[4]	Phase	Parent Study Accession
AA_CAC	phs001412	NHLBI TOPMed: African American Coronary Artery Calcification (AA CAC)	DHS	Allred	339	BROAD	2	
AA_CAC, GeneSTAR	phs001218	NHLBI TOPMed: GeneSTAR (Genetic Study of Atherosclerosis Risk)	GeneSTAR	Mathias	1,639	MACROGEN, BROAD[5]	2	phs001074
AA_CAC, HyperGEN_GENOA	phs001345	NHLBI TOPMed: Genetic Epidemiology Network of Arteriopathy (GENOA)	GENOA	Peyser & Kardia	1,143	BROAD, UW	2	phs001238
AA_CAC, MESA	phs001416	NHLBI TOPMed: MESA and MESA Family AA-CAC	MESA	Rich & Rotter	4,819	BROAD	2	phs000209
AFGen	phs000997	NHLBI TOPMed: The Vanderbilt AF Ablation Registry	VAFAR	Shoemaker	154	BROAD	1	
AFGen	phs001024	NHLBI TOPMed: Partners HealthCare Biobank	Partners	Lubitz	111	BROAD	1	
AFGen	phs001032	NHLBI TOPMed: The Vanderbilt Atrial Fibrillation Registry	VU_AF	Darbar	1,018	BROAD	1	
AFGen	phs001040	NHLBI TOPMed: Novel Risk Factors for the Development of Atrial Fibrillation in Women	WGHS	Albert & Chasman	98	BROAD	1	
AFGen	phs001062	NHLBI TOPMed: MGH Atrial Fibrillation Study	MGH_AF	Lubitz	918	BROAD	1	phs001001
AFGen	phs001189	NHLBI TOPMed: Cleveland Clinic Atrial Fibrillation Study	CCAF	Chung & Barnard	329	BROAD	1	phs000820
AFGen, FHS	phs000974	NHLBI TOPMed: Whole Genome Sequencing and Related Phenotypes in the Framingham Heart Study	FHS	Ramachandran	3,749	BROAD	1	phs000007
Amish	phs000956	NHLBI TOPMed: Genetics of Cardiometabolic Health in the Amish	Amish	Mitchell	1,028	BROAD	1	
BAGS	phs001143	NHLBI TOPMed: The Genetics and Epidemiology of Asthma in Barbados	BAGS	Barnes	962	ILLUMINA	1	
CFS	phs000954	NHLBI TOPMed: The Cleveland Family Study (WGS)	CFS	Redline	920	UW	1	phs000284
COPD	phs000946	NHLBI TOPMed: Boston Early-Onset COPD Study in the TOPMed Program	EOCOPD	Silverman	66	UW	1	phs001161
COPD	phs000951	NHLBI TOPMed: Genetic Epidemiology of COPD (COPDGene) in the TOPMed Program	COPDGene	Silverman	8,742	BROAD, UW	1, 2	phs000179
CRA_CAMP	phs000988	NHLBI TOPMed: The Genetic Epidemiology of Asthma in Costa Rica	CRA	Weiss	1,041	UW	1	
GenSalt	phs001217	NHLBI TOPMed: Genetic Epidemiology Network of Salt Sensitivity (GenSalt)	GenSalt	He	1,695	BAYLOR	2	phs000784
GOLDN	phs001359	NHLBI TOPMed: Genetics of Lipid Lowering Drugs and Diet Network (GOLDN)	GOLDN	Arnett	904	UW	2	phs000741
HyperGEN_GENOA	phs001293	NHLBI TOPMed: HyperGEN - Genetics of Left Ventricular (LV) Hypertrophy	HyperGEN	Arnett	1,776	UW	2	
JHS	phs000964	NHLBI TOPMed: The Jackson Heart Study	JHS	Correa	3,128	UW	1	phs000286
PGX_Asthma	phs000920	NHLBI TOPMed: Genes-environments and Admixture in Latino Asthmatics (GALA II) Study	GALAII	Burchard	913	NYGC[5]	1	phs001180
PGX_Asthma	phs000921	NHLBI TOPMed: Study of African Americans, Asthma, Genes and Environment (SAGE) Study	SAGE	Burchard	451	NYGC[5]	1	
SAFS	phs001215	NHLBI TOPMed: San Antonio Family Heart Study (WGS)	SAFS	Blangero	1,502	ILLUMINA	1	
Sarcoidosis	phs001207	NHLBI TOPMed: African American Sarcoidosis Genetics Resource	Sarcoidosis	Montgomery	608	BAYLOR	2	
SAS	phs000972	NHLBI TOPMed: Genome-wide Association Study of Adiposity in Samoans	SAS	McGarvey	1,208	UW, NYGC	1, 2	phs000914
THRV	phs001387	Rare Variants for Hypertension in Taiwan Chinese (THRV)	THRV	Rao & Chen	1,525	BAYLOR	2	
VTE	phs001368	NHLBI TOPMed: Cardiovascular Health Study	CHS	Heckbert	69	BAYLOR	2	phs000287
VTE	phs001402	NHLBI TOPMed: Whole Genome Sequencing of Venous Thromboembolism (WGS of VTE)	Mayo_VTE	de Andrade	1,251	BAYLOR	2	phs000289
VTE, AFGen	phs000993	NHLBI TOPMed: Heart and Vascular Health Study (HVH)	HVH	Heckbert & Smith	614	BROAD, BAYLOR	1, 2	phs001013
VTE, AFGen	phs001211	NHLBI TOPMed: Trans-Omics for Precision Medicine Whole Genome Sequencing Project: ARIC	ARIC	Boerwinkle	3,612	BAYLOR, BROAD	1, 2	phs000280

WHI	phs001237	NHLBI TOPMed: Women's Health Initiative (WHI)	WHI	Kooperberg	10,047	BROAD	2	phs000200
			TOTAL SAMPLES		**54,854**			

1 - TOPMed Project. AFGen=Atrial Fibrillation Genetics Consortium. Amish=Genetics of Cardiometabolic Health in the Amish; BAGS=Barbados Asthma Genetics Study; CFS=Cleveland Family Study; COPD=Genetic Epidemiology of COPD; CRA_CAMP=The Genetic Epidemiology of Asthma in Costa Rica and the Childhood Asthma Management Program; FHS=Framingham Heart Study; JHS=Jackson Heart Study; PGX_Asthma=Pharmacogenomics of Bronchodilator Response in Minority Children with Asthma; SAS=Samoan Adiposity Study; VTE=Venous Thromboembolism; AA_CAC=African American Coronary Artery Calcification; GeneSTAR=Genetic Studies of Atherosclerosis Risk; GenSalt=Genetic Epidemiology Network of Salt Sensitivity; GOLDN=Genetics of Lipid Lowering Drugs and Diet Network; HyperGEN_GENOA=Hypertension Genetic Epidemiology Network and Genetic Epidemiology Network of Arteriopathy; MESA=Multi-Ethnic Study of Atherosclerosis; SAFS=San Antonio Family Studies; Sarcoidosis=Genetics of Sarcoidosis in African Americans; WHI=Women's Health Initiative. Project descriptions are available on the TOPMed website, https://topmed.nhlbi.nih.gov [20].

2 - Study name as it appears in dbGaP

3 - Approximate sample size for freeze5b release

4 - NYGC = New York Genome Center; BROAD = Broad Institute of MIT and Harvard; UW = University of Washington Northwest Genomics Center; ILLUMINA = Illumina Genomic Services; MACROGEN = Macrogen Corp.; BAYLOR = Baylor Human Genome Sequencing Center

5 - ILLUMINA was an additional sequencing center for legacy data contributed by GALAII (n=6 samples), SAGE (n=10 samples) and GeneSTAR (n=283 samples).

Please note that most (but not all) samples in previous releases (genotype call sets for Freezes 3 and 4) are included in Freeze 5b (along with many new samples). Because some investigators are in the process of analyzing Freeze 4 data across multiple studies and because it includes some samples that are not in Freeze 5b, the Freeze 4 call set will also be included along with Freeze 5b in the 2018 dbGaP releases.

The following sections of this document describe methods of data acquisition, processing and quality control (QC) for TOPMed WGS data contained in the 2018 Freeze 5b call set. (A separate document describes methods for the Freeze 4 call set.) Briefly, ~30X whole genome sequencing was performed at several different Sequencing Centers (named in Table 1). In most cases, all samples for a given study were sequenced at the same center (see Table 1 for exceptions), except for a small number of control samples described below. The reads were aligned to human genome build GRCh37 or GRCh38 at each center using similar, but not identical, processing pipelines. The resulting sequence data files were transferred from all centers to the TOPMed Informatics Research Center (IRC), where they were re-aligned to build GRCh38, using a common pipeline to produce a set of 'harmonized' .cram files. The IRC performed joint genotype calling on all samples in the Freeze 5b release (as well as additional studies to be released at a later time). The resulting VCF files were split by study and consent group for distribution to approved dbGaP users. They can be reassembled easily for cross-study, pooled analysis since the files for all studies contain identical lists of variant sites. Quality control was performed at each stage of the process by the Sequencing Centers, the IRC and the TOPMed Data Coordinating Center (DCC). Only samples that passed QC are included in the call set, whereas all variants (whether passed or failed) are included.

Genotype call sets are provided in VCF format, with one file per chromosome. GRCh38 read alignments are not provided currently by dbGaP, but there is a plan to do so in the future.

TOPMed DNA sample/sequencing-instance identifiers

Each DNA sample processed by TOPMed is given a unique identifier as "NWD" followed by six digits (e.g. NWD123456). These identifiers are unique across all TOPMed studies. Each NWD identifier is associated with a single study subject identifier used in other dbGaP files (such as phenotypes, pedigrees and consent files). A given subject identifier may link to multiple NWD identifiers if duplicate samples are sequenced from the same individual. Study investigators assign NWD IDs to subjects. Their biorepositories assign DNA samples and NWD IDs to specific bar-coded wells/tubes supplied by the Sequencing Center and record those assignments in a sample manifest, along with other metadata (e.g. sex, DNA extraction method). At each Sequencing Center, the NWD ID is propagated through all phases of the pipeline and is the primary identifier in all results files. Each NWD ID results in a single sequencing instance and is linked to a single subject identifier in the sample-subject mapping file for each accession. In contrast to the project wide NWD identifiers, subject identifiers are study-specific and may not be unique across all of TOPMed accessions.

Control Samples

In Phase 1, one parent-offspring trio from the Framingham Heart Study (FHS) was sequenced at each of four Sequencing Centers (family ID 746, subject IDs 13823, 15960 and 20156) All four WGS runs for each subject are provided in the TOPMed FHS accession (phs000974). In Phase 2, one 1000G Puerto Rican Trio (HG01110, HG01111, HG01249) was sequenced once at each center. In addition, HapMap subjects NA12878 (CEU, Lot K6) and NA19238 (YRI, Lot E2) were sequenced at each of the Sequencing Centers in alternation, once approximately every 1000 study samples throughout both Phases 1 and 2. The 1000G and HapMap sequence data will be released publicly as a BioProject in the future.

Sequencing Center Methods

Broad Institute of MIT and Harvard

Stacey Gabriel

The Broad sequenced several studies in each of Phases 1 and 2 (see Table 1). The methods described below showcase the process for phase 1 and highlight the changes that were implemented during phase 2.

DNA Sample Handling and QC

DNA samples were informatically received into the Genomics Platform's Laboratory Information Management System via a scan of the tube barcodes using a Biosero flatbed scanner. This registered the samples and enabled the linking of metadata based on well position. All samples were then weighed on a BioMicro Lab's XL20 to determine the volume of DNA present in sample tubes. For some of the latter phase 2 samples, the DNA volume measurements were performed using Dynamic Devices' Lynx VVP since this switch was made in production at large. Following this, the samples were quantified in a process that uses PICO-green fluorescent dye. Once volumes and concentrations were determined, the samples were handed off to the Sample Retrieval and Storage Team for storage in a locked and monitored -20 walk-in freezer.

Library Construction

Samples were fragmented by means of acoustic shearing using Covaris focused-ultrasonicator, targeting 385 bp fragments. Following fragmentation, additional size selection was performed using a SPRI cleanup. Library preparation was performed using a commercially available kit provided by KAPA Biosystems (product KK8202) with palindromic forked adapters with unique 8 base index sequences embedded within the adapter (purchased from IDT). Following sample preparation, libraries were quantified using quantitative PCR (kit purchased from KAPA biosystems) with probes specific to the ends of the adapters. This assay was automated using Agilent's Bravo liquid handling platform and run on a ViiA 7 from Thermo Fisher. Based on qPCR quantification, libraries were normalized to 1.7 nM. For the majority of Phase 1, samples were pooled into 8-plexes and the pools were once again qPCRed, and normalized to 1.2nM. For the end of Phase 1, and all of Phase 2, samples were pooled in 24-plexes. Samples were then combined with HiSeq X Cluster Amp Mix 1,2 and 3 into single wells on a strip tube using the Hamilton Starlet Liquid Handling system.

Clustering and Sequencing

Both TOPMed phase 1 and phase 2 followed the same process except for version changes in the softwares. As described in the library construction process, 96 samples on a plate were processed together through library construction. A set of 96 barcodes was used to index the samples. Barcoding allows pooling of samples prior to loading on sequencers and mitigates lane-lane effects at a single sample level. For the beginning of Phase 1, the plate was broken up into 12 pools of 8 samples each, and for the end of Phase 1 and all of Phase 2, the plate was broken up into 4 pools of 24 samples each. For 8-plex pooling, pools were taken as columns on the plate (e.g., each column comprises a pool). From this format (and given the current yields of a HiSeqX) each pool was then spread over 1 flowcell (8 lanes). For 24 plex pooling, the four pools were taken as columns on the plate (e.g., columns 1-3; 4-6; 7-9; 10-12). From this format (and given the current yields of a HiSeqX) the 4 pools were spread over 3 flowcells (24 lanes). Cluster amplification of the templates was performed according to the manufacturer's protocol (Illumina) using the Illumina cBot. For phase 1, flowcells were sequenced on Hi Seq X with sequencing software HiSeq Control Software (HCS) versions 3.1.26 and 3.3.39, then analyzed using RTA2 (Real Time Analysis) versions 2.3.9 and 2.7.1. For phase 2, the versions of the sequencing software used were HiSeq Control Software (HCS) versions 3.3.39, 3.3.76 and HD 3.4.0.38, and then analyzed using RTA2 versions 2.7.1, 2.7.6, and 2.7.7. During all of Phase 1, sequencing was done with only reading a single index. To mitigate the "index hopping" phenomenon, dual index reads was incorporated in the middle of Phase 2.

Read Processing

For TOPMED phase 1 data, the following versions were used for aggregation, and alignment to Homo_sapiens_assembly19_1000genomes_decoy reference: picard (latest version available at the time of the analysis), GATK (3.1-144-g00f68a3) and BwaMem (0.7.7-r441).

For TOPMED phase 2 data, we used the following versions for the on-prem data generation for aggregation, and alignment to Homo_sapiens_assembly19_1000genomes_decoy reference or Homo_sapiens_assembly19: picard (latest version available at the time of the analysis), GATK (3.1-144-g00f68a3) and BwaMem (0.7.7-r441). For the data that were generated on the cloud as part of Phase 2, we used the following versions for aggregation, and alignment to Homo_sapiens_assembly38: picard (latest version available at the time of analysis, ranges from 1.1150 - 2.12.0), BQSR: latest available (GATK 4.alpha-249-g7df4044 - 4.beta.5) and BwaMem: 0.7.15.r1140.

Sequence Data QC

A sample was considered sequence complete when the mean coverage was >= 30x. Two QC metrics that were reviewed along with the coverage are the sample Fingerprint LOD score (score which estimates the probability that the data is from a given individual, see below for more details) and % contamination. At aggregation, an all-by-all comparison of the read group data and estimation of the likelihood that each pair of read groups is from the same individual were performed. If any pair had a LOD score < -20, the aggregation did not proceed and was investigated. FP LOD >= 3 was considered passing concordance with the sequence data (ideally LOD >10). A sample will have an LOD of 0 when the sample failed to have a passing fingerprint. Fluidigm fingerprint was repeated once if failed. Read groups with fingerprint LODs of < -3 were blacklisted from the aggregation. If the sample did not meet coverage, it was topped off for additional coverage. If a large % of read groups were blacklisted, it was investigated as a potential sample swap. In terms of contamination, a sample was considered passing if the contamination was less than 3%. In general, the bulk of the samples had less than 1% contamination.

Fingerprinting

For the purpose of fingerprinting we extract a small aliquot from each sample prior to any of the processing for sequencing. This aliquot is genotyped on a set of 96 common SNPs. These SNPs have been carefully selected so that they are enable the identity validation of each of our read groups separately. This ensures that the aggregated sample (comprising of about 24 reads groups) consist of data only from the intended sample. The genotyping is performed using a Fluidigm AccessArray with our custom SNPs and the comparison is done using Picard's CheckFingerprints which calculates the LogOddsRatio (LOD) of the sequence data matching versus not matching the genotype data.

Northwest Genomics Center

Deborah Nickerson

The NWGC performed sequencing on several studies from each of Phases 1 and 2. The methods given below were the same for Phases 1 and 2, except where noted otherwise. For Phase 1, all samples were sequenced at Macrogen (with methods described in this section); for Phase 2, some samples were sequenced at Macrogen and others at NWGC.

DNA Sample Handling and QC

The NWGC centralized all receipt, tracking, and quality control/assurance of DNA samples in a Laboratory

Information Management System. Samples were assigned unique barcode tracking numbers and had a detailed sample manifest (i.e., identification number/code, sex, DNA concentration, barcode, extraction method). Initial QC entailed DNA quantification, sex typing, and molecular "fingerprinting" using a high frequency, cosmopolitan genotyping assay. This 'fingerprint' was used to identify potential sample handling errors and provided a unique genetic ID for each sample, which eliminated the possibility of sample assignment errors. In addition, ~8% of the samples per batch were spot checked on an agarose gel to check for high molecular weight DNA; if DNA degradation was detected all samples were checked. Samples were failed if: (1) the total amount, concentration, or integrity of DNA was too low; (2) the fingerprint assay produced poor genotype data or (3) sex-typing was inconsistent with the sample manifest. Barcoded plates were shipped to Macrogen for library construction and sequencing.

Library Construction

Libraries were constructed with a minimum of 0.4ug gDNA and were prepared in Covaris 96 microTUBE plates and sheared through a Covaris LE220 focused ultrasonicator targeting 350 bp inserts. The resulting sheared DNA was selectively purified using sample purification beads to make the precise length of insert; End-repair (repaired to blunt end), A-tailing (A-base is added to 3'end), and ligation (Y-shaped adapter is used which includes a barcode) were performed as directed by TruSeq PCR-free Kit (Illumina, cat# FC-121-3003) protocols for Phase 1 studies, and by KAPA Hyper Prep Kit without amplification (KR0961.v1.14) for Phase 2 studies. A second Bead cleanup was performed after ligation to remove any residual reagents and adapter dimers. To verify the size of adapter-ligated fragments, the template size distribution was validated by running on a 2200 TapeStation (Agilent, Catalog # G2964AA) using a TapeStation DNA Screen Tape (Agilent, Catalog 5067-5588). The final libraries were quantified by qPCR assay using KAPA library quantification kit (cat.# KK4808 and KK4953) on a Light Cycler 480 instrument (Roche, cat# 05015278001).

Clustering and Sequencing

Eight normalized and indexed libraries were pooled together and denatured before cluster generation on a cBot. The 8-plex pools were loaded on eight lanes of a flow cell and sequenced on a HiSeqX using illumina's HiSeq X ten reagents kit (V2.5, cat# FC-501-2521). For cluster generation, every step was controlled by cBot. When cluster generation was complete, the clustered patterned flow cells were then sequenced with sequencing software HCS (HiSeq Control Software). The runs were monitored for %Q30 bases using the SAV (Sequencing Analysis Viewer). Using RTA 2 (Real Time Analysis 2) the BCLs (base calls) were de-multiplexed into individual FASTQs per sample using illumina package bcl2fastq v2.15.0 and transferred from Macrogen to NWGC for alignment, merging, variant calling and sequencing QC.

Read Processing

The processing pipeline consisted of aligning FASTQ files to a human reference (hs37d5;1000 Genomes hs37d5 build 37 decoy reference sequence) using BWA-MEM (Burrows-Wheeler Aligner; v0.7.10) (Li and Durbin 2009). All aligned read data were subject to the following steps: (1) "duplicate removal" was performed, (i.e., the removal of reads with duplicate start positions; Picard MarkDuplicates; v1.111) (2) indel realignment was performed (GATK IndelRealigner; v3.2) resulting in improved base placement and lower false variant calls, and (3) base qualities were recalibrated (GATK BaseRecalibrator; v3.2). Sample BAM files were "squeezed" using Bamutil with default parameters and checksummed before being transferred to the IRC.

Sequence Data QC

All sequence data underwent a QC protocol before being released to the TOPMed IRC for further processing. For whole genomes, this included an assessment of: (1) mean coverage; (2) fraction of genome covered greater than 10x; (3) duplicate rate; (4) mean insert size; (5) contamination ratio; (6) mean Q20 base coverage; (7) Transition/Transversion ratio (Ti/Tv); (8) fingerprint concordance > 99%; and (9) sample homozygosity and heterozygosity. All QC metrics for both single-lane and merged data were reviewed by a sequence data analyst to identify data deviations from known or historical norms. Lanes/samples that failed QC were flagged in the system and were re-queued for library prep (< 1% failure) or further sequencing (< 2% failure), depending upon the QC issue.

New York Genome Center

Soren Germer

The NYGC performed sequencing for several studies in each of Phases 1 and 2 (see Table 1). The methods were the same for Phases 1 and 2, except where noted otherwise.

DNA Sample Handling and QC

Genomic DNA samples were submitted in NYGC-provided 2D barcoded matrix rack tubes. Sample submissions were randomized either at investigator laboratory or upon receipt at NYGC (using a BioMicroLab XL20). Upon receipt, the matrix racks were inspected for damage and scanned using a VolumeCheck instrument (BioMicroLab), and tube barcode and metadata from the sample manifest uploaded to NYGC LIMS. Genomic DNA was quantified using the Quant-iT PicoGreen dsDNA assay (Life Technologies) on a Spectramax fluorometer, and the integrity was ascertained on a Fragment Analyzer (Advanced Analytical). After sample quantification, a separate aliquot (100ng) was removed for SNP array genotyping with the HumanCoreExome-24 array (Illumina). Array genotypes were used to estimate sample contamination (using VerifyIDintensity), for sample fingerprinting, and for downstream quality control of sequencing data. Investigator was notified of samples that failed QC for total mass, degradation or contamination, and replacement samples were submitted.

Library Construction

Sequencing libraries were prepared with 500 ng DNA input, using the TruSeq PCR-free DNA HT Library Preparation Kit (Illumina) for Phase 1 samples and the Kappa Hyper Library Preparation Kit (PCR-free), following manufacturer's protocol with minor modifications to account for automation. Briefly, genomic DNA was sheared using the Covaris LE220 sonicator to a target size of 450 bp (t:78; Duty:15; PIP:450; 200 cycles), followed by end-repair and bead based size selection of fragmented molecules (0.8X). The selected fragments were A-tailed, and sequence adaptors ligated onto the fragments, followed by two bead clean-ups of the libraries (0.8X). These steps were carried out on the Caliper SciClone NGSx workstation (Perkin Elmer). Final libraries are evaluated for size distribution on the Fragment Analyzer or BioAnalyzer and quantified by qPCR with adaptor specific primers (Kapa Biosystems).

Clustering and Sequencing

Final libraries were multiplexed for 8 samples per sequencing lane, with each sample pool sequenced across 8 flow cell lanes. 1% PhiX control was spiked into each library pool. The library pools were quantified by qPCR, loaded on the to HiSeq X patterned flow cells and clustered on an Illumina cBot following manufacturer's protocol. Flow cells were sequenced on the Illumina HiSeq X with 2x150bp reads, using V2 (Phase 1) or V3 (Phase 2) sequencing chemistry, and Illumina HiSeq Control Software v3.1.26 (Phase 1) or HCS3.3.39 (Phase 2).

Read Processing

Demultiplexing of sequencing data was performed with bcl2fastq2 (v2.16.0.10 for Phase 1 and v2.17.1.14 for Phase 2), and sequencing data was aligned to human reference build 37 (hs37d5 with decoy) using BWA-MEM (v0.7.8 for Phase 1 and v0.7.12 for Phase 2). Data was further processed using the GATK best-practices pipeline (v3.2-2 fo Phase 1 and v3.4-0 for Phase 2), with duplicate marking using Picard tools (v1.83 for Phase 1 and v1.137 for Phase 2), realignment around indels, and base quality recalibration. Individual sample BAM files were squeezed using Bamutil v1.0.9 with default parameters -- removing OQ's, retaining duplicate marking and binning quality scores (binMid) -- and transferred to the IRC using Globus. Individual sample SNV and indel calls were generated using GATK haplotype caller and joint genotyping was performed across all the NYGC phase 1 samples.

Sequence Data QC

Prior to release of BAM files to IRC, we ensured that mean genome coverage was >=30x, when aligning to the ~2.86Gb sex specific mappable genome, and that uniformity of coverage was acceptable (>90% of genome covered >20x). Sample identity and sequencing data quality was confirmed by concordance to SNP array genotypes. Sample contamination was estimated with VerifyBAMId v1.1.0 (threshold <3%). Gender was determined from X- and Y-chromosome coverage and checked against submitter information. Further QC included review of alignment rates, duplicate rates, and insert size distribution. Metrics used for review of SNV and indel calls included: the total number of variants called, the ratio of novel to known variants, and the Transition to Transversion ratios, and the ratio of heterozygous to homozygous variant calls.

Illumina Genomic Services

Karine Viaud Martinez

Two Phase 1 studies were sequenced by Illumina Genomic Services: BAGS (phs001143) and SAFS (phs001215). Methods were the same for both studies, except for those in the "Clustering and Sequencing" section below. Additional studies have provided small numbers of "legacy" samples. These were sequenced by Illumina to 30x depth prior to the start of the TOPMed project and have been remapped and included in the freeze 5b call set.

DNA Sample Handling and QC

Project samples were processed from 96-well barcoded plates provided by Illumina. Electronic manifest including unique DNA identification number describing the plate barcode and well position (eg, LP6002511-DNA_A01) and samples information (e.g. Gender, Concentration, Volume, Tumor/normal, Tissue type, Replicate…) was accessioned in LIMS. This enabled a seamless interface with robotic processes and retained sample anonymity. An aliquot of each sample was processed in parallel through the Infinium Omni 2.5M (InfiniumOmni2.5Exome-8v1, HumanOmni25M-8v1) genotyping array and an identity check was performed between the sequencing and array data via an internal pipeline. Genomic DNA was quantified prior to library construction using PicoGreen (Quant-iT™ PicoGreen® dsDNA Reagent, Invitrogen, Catalog #: P11496). Quants were read with Spectromax Gemini XPS (Molecular Devices).

Library Construction

Samples were batched using LIMS, and liquid handling robots performed library preparation to guarantee accuracy and enable scalability. All sample and reagent barcodes were verified and recorded in LIMS. Paired-end libraries were generated from 500ng-1ug of gDNA using the Illumina TruSeq DNA Sample Preparation Kit (Catalog #: FC-121-2001), based on the protocol in the TruSeq DNA PCR-Free Sample Preparation Guide. Pre-fragmentation gDNA cleanup was performed using paramagnetic sample purification beads (Agencourt® AMPure® XP reagents, Beckman Coulter). Samples were fragmented and libraries are size selected following fragmentation and end-repair using paramagnetic sample purification beads, targeting short insert sizes. Final libraries were quality controlled for size using a gel electrophoretic separation system and awee quantified.

Clustering and Sequencing

BAGS (phs001143) study: Following library quantitation, DNA libraries were denatured, diluted, and clustered onto v4 flow cells using the Illumina cBot™ system. A phiX control library was added at approximately 1% of total loading content to facilitate monitoring of run quality. cBot runs were performed based on the cBot User Guide, using the reagents provided in Illumina TruSeq Cluster Kit v4. Clustered v4 flow cells were loaded onto HiSeq 2000 instruments and sequenced on 125 bp paired-end, non-indexed runs. All samples were sequenced on independent lanes. Sequencing runs were performed based on the HiSeq 2000 User Guide, using Illumina TruSeq SBS v4 Reagents. Illumina HiSeq Control Software (HCS) and Real-Time Analysis (RTA) were used on HiSeq 2000 sequencing runs for real-time image analysis and base calling.

SAFS (phs 001215) study : Following library quantitation, DNA libraries were denatured, diluted and clustered onto patterned flow cells using the Illumina cBot™ system. A phiX control library was added at approximately 1% of total loading content to facilitate monitoring of run quality. cBot runs were performed following cBot System Guide, using Illumina HiSeq X HD Paired End Cluster Kit reagents. Clustered patterned flow cells were loaded onto HiSeq X instruments and sequenced on 151 bp paired-end, non-indexed runs. All samples were sequenced on independent lanes. Sequencing runs were performed based on the HiSeq X System Guide, using HiSeq X HD SBS Kit reagents. Illumina HiSeq Control Software (HCS), and Real-Time Analysis (RTA) were used with the HiSeq X© sequencers for real-time image analysis, and base calling.

Read Processing

The Whole Genome Sequencing Service leverages a suite of proven algorithms to detect genomic variants comprehensively and accurately. Most versions of the Illumina callers are open source and available publicly. See the Illumina GitHub (https://github.com/Illumina [21]) for the current releases. One or more lanes of data were processed from run folders directly with the internal use only ISAS framework (2.5.55.16 or 2.5.26.13 depending on the start of the project), including alignment with iSAAC (iSAAC-01.14.02.06 or iSAAC-SAAC00776.15.01.27), small variants called with Starling (2.0.17 or starka-2.1.4.2), structural variants called with Manta (manta-0.18.1 or manta-0.23.1) and copy number variants with Canvas (v4.0).

Sequence Data QC

The genome build QC pipeline was automated to evaluate both primary (sequencing level) and secondary (build level) metrics against expectations based on historical performance. Multiple variables, such as Gb of high quality (Q30) data, mismatch rates, percentage of aligned reads, insert size distribution, concordance to the genotyping array run in parallel, average depth of coverage, number of variants called, callability of the genome as a whole as well as of specific regions (evenness of coverage), het/hom ratio, duplicate rates, and noise were assessed. Genome builds that were flagged as outliers at QC are reviewed by our scientists for investigation. Scientists reviewed all QC steps during the process: Library quantification and fragment size; run quality; genotyping and sequencing data considering Sample Manifest information (Tumor/Normal, tissue type). Libraries or sequencing lanes were requeued for additional sequencing or library prep as needed.

Macrogen

Sal Situ

In collaboration with NWGC, Macrogen participated in the sequencing of several Phase 1 studies, as described above. In addition, Macrogen independently performed sequencing of one Phase 2 study, GeneSTAR (phs001218), using the following methods.

DNA Sample Handling and QC

Macrogen centralized all receipt, tracking, and quality control/assurance of DNA samples in a Laboratory Information Management System (LIMS). Samples had a detailed sample manifest (i.e., identification number/code, sex, DNA concentration, barcode, extraction method). Initial QC entailed DNA quantification using Quant-iT PicoGreen dsDNA assay (Life Technologies, cat# P7589).

Library Construction

Starting with minimum of 0.4 ug of DNA, samples were sheared in a 96-well format using a Covaris LE220 focused ultrasonicator targeting 350 bp inserts. The resulting sheared DNA was selectively purified by sample purification beads to make the precise length of insert. End-repair, A-tailing, and ligation were performed as directed by KAPA Hyper Prep Kit(KAPA Biosystems, cat.# KK8505) without amplification (KR0961 v1.14) protocols. A second Bead cleanup was performed after ligation to remove any residual reagents and adapter dimers.

Clustering and Sequencing

Prior to sequencing, final library concentration was determined by duplicate qPCR using the KAPA Library Quantification Kit (KK4854), and molecular weight distributions were verified using the TapeStation2200. Samples were sequenced on a HiSeq X using Illumina's HiSeq X Ten Reagent Kit (v2.5) with 2*150bp reads. Briefly, validated libraries were denatured, diluted and clustered onto v2.5 flow cells using the Illumina cBot system. The clustered patterned flow cells were then sequenced with sequencing software HCS (HiSeq Control Software, version 3.5.0.7). The runs were monitored for %Q30 bases and %PF reads using the SAV (Sequencing Analysis Viewer version 1.10.2).

Read Processing

Illumina sequencing instruments, including HiSeqX, generate per-cycle BCL base call files as primary sequencing output. These BCL files were aligned with ISAAC (v.01.15.02.08) to GRCh37/hg19 from UCSC.

Before aligning steps, the proportion of base quality (Q30) was checked. If Q30 < 80%, the sample was re-sequenced. During alignment steps, the duplicated reads were marked and not used for variant calling.

For the downstream analysis applications, we also provided FASTQ files via bcl2fastq software (v. 2.17).

Sequence Data QC

After finishing alignment, the overall QC was conducted and a sample passed if,

1) the mappable mean depth is higher than 30X

2) the proportion of regions covered more than 10X is greater than 95%

3) contamination rates (Freemix: ASN, EUR) are less than 3% determined by VerifyBamID.

Moreover, we check the proportion of GC, insert size, and Depth of Coverage (mode of sequence depth, interquartile range of depth and distance from Poisson distribution), when the proportion of 10X coverage failed.

Baylor College of Medicine Human Genome Sequencing Center

Richard Gibbs

The Baylor HGSC sequenced several Phase 2 studies (see Table 1), using the following methods.

DNA Sample Handling and QC

Once samples were received at the HGSC, sample tube barcodes were scanned into the HGSC LIMS using a flatbed barcode scanner and marked as 'received' by the sample intake group. The sample number and barcodes relative to rack position were checked and any physical discrepancies and/or inconsistencies with respect to the sample manifest were noted and reported. The approved sample manifest containing the designated metadata was then directly uploaded into the HGSC LIMS. The metadata were linked at intake to a unique and de-identified sample identifier (NWD ID), which was propagated through all phases of the pipeline. This unique identifier was subsequently embedded in the library name, all sequencing events, and all deliverable files.

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. This assay was setup in 384-well plates using a Biomek 2000 robot and fluorescence determined using the Synergy 2 fluorescence spectrophotometer. Semi-quantitative and qualitative "yield gels" were used to estimate DNA sample integrity. DNA was run on 1 % E-gels (Life Tech Inc.) along with known and DNA standards previously used in the Picogreen assay and 1 Kb (NEB) DNA size ladder. These gels also served indirectly as a "cross-validation" for the Picogreen assay since the same standards were used in both assays. To ensure sample identity and integrity, an orthogonal SNP confirmation was used for the TOPMed samples by employing a panel of 96 SNP loci selected by the Rutgers University Cell and DNA Repository (RUCDR). This assay addresses specific attributes around gender, and polymorphisms across populations and ancestry. This panel of 96 SNP loci is commercially available through Fluidigm as the SNPtrace™ Panel. The workflow includes Fluidigm Integrated Fluidic Circuits (IFCs) that utilizes the allele-specific PCR-based Fluidigm SNPtype assay to process 9216 genotypes (96 sites x 96 samples). This SNP panel serves the QA/QC process by distinguishing closely related samples and duplicate samples, and verifying gender with the reported manifest value prior to sequencing. It also assists in early stage contamination detection, and is used to validate sample concordance against the final sequence files to ensure pipeline integrity.

Library Construction

Libraries were routinely prepared using Beckman robotic workstations (Biomek FX and FXp models) in batches of 96 samples and all liquid handling steps were incorporated into the LIMS tracking system. To ensure even coverage of the genome, KAPA Hyper PCR-free library reagents (KK8505, KAPA Biosystems Inc.) were used for library construction. DNA (500 ng) was sheared into fragments of approximately 200-600 bp in a Covaris E220 system (96 well format) followed by purification of the fragmented DNA using AMPure XP beads. A double size selection step was then employed, with different ratios of AMPure XP beads, to select a narrow band of sheared DNA for library preparation. DNA end-repair and 3'-adenylation were performed in the same reaction followed by ligation of the barcoded adaptors to create PCR-Free libraries. The Fragment Analyzer (Advanced Analytical Technologies, Inc.) instrument was used to assess library size and presence of remaining adapter dimers. This protocol allowed for the routine preparation of 96-well library plates in 7 hours. For Library size estimation and quantification, the library was run on Fragment Analyzer (Advanced Analytical Technologies, Inc., Ames, Iowa) followed by qPCR assay using KAPA Library Quantification Kit using their SYBR® FAST qPCR Master Mix. Both of these assays were done in batches of 96 samples in 3-4 hours. Automated library construction and quantification procedures routinely included a positive and negative control (no template control) on every 96-well library construction plate to monitor process consistency and possible contamination events. Standard library controls utilized NA12878 (NIST Gold Standard Hapmap sample) as the primary comparison sample. In accordance with TOPMed protocols we also included control standard supplied by the TOPMed program in every 10th plate of processed libraries.

Clustering and Sequencing

WGS libraries were sequenced on the Illumina HiSeq X Ten instrument fleet to generate 150 bp paired-end sequence. Optimal library concentrations used for cluster generation were determined before releasing libraries into production. Typical loading concentrations range between 150-450pM._Libraries were loaded using the HiSeq X(tm) Ten Reagent Kit v2.5. Run performance was monitored through key metrics using the current HiSeq X instrument software (3.3.39) to assess cluster density, signal intensity and phasing/pre-phasing. One sample was loaded per HiSeq X lane to achieve a minimum coverage of 30X, or 90 Gbp of unique reads aligning to the human reference per sample. Each of these metrics was evaluated to confirm library quality and concentration and to detect any potential chemistry, reagent delivery and/or optical issues. Overall run performance was evaluated by metrics from the off-instrument software (Casava) and from mapping results generated by the Mercury (HgV) analysis pipelines.

Read Processing

All sequencing events were subject to the HgV Human Resequencing Protocol, which included BCL conversion to FASTQ, BWA-MEM mapping, GATK recalibration and realignment. All multiplexed flow cell data (BCLs) were converted to barcoded FASTQs, which were aligned via BWA-MEM to the GRCh37 decoy reference. The resulting sequence event (SE) BAMs were assessed for barcode, lane, and flow cell QC metrics including contamination (VerifyBamID) using a set of HapMap-derived MAFs. Duplicate, unmapped, and low quality reads were flagged rather than filtered. Sample BAMs were then GATK-recalibrated and realigned using dbSNP142b37, 1KGP Phase 1 and Mills gold standard indels. BAM files were "squeezed" by stripping multiple tags and binning the quality scores, resulting in the final deliverable of a ~60 GB BAM.

Sequence Data QC

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 90 GB unique aligned bases per lane. Genome coverage metrics were also tracked to achieve 90% of genome covered at 20x and 95% at 10x with a minimum of 86 x 10^9 mapped, aligned bases with Q20 or higher. 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. 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 SNP Trace 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.

Informatics Research Center Methods

Tom Blackwell, Hyun Min Kang and Gonçalo Abecasis

Center for Statistical Genetics, Department of Biostatistics, University of Michigan

The IRC pipeline consists of two major processes diagrammed in the Figure 1 below: (1) Harmonization of data from the BAM files provided by the Sequencing Centers and (2) joint variant discovery and genotype calling across studies. Detailed protocols for these processes are given in the following sections.

Figure 1 : Schematic view of IRC alignment and variant calling pipeline

Harmonization of Read Alignments

Ahead of joint variant discovery and genotype calling by the IRC, the sequence data obtained from the TOPMed Sequencing Centers were remapped using a standard protocol to produce "harmonized" sequence data files.

Sequence data were received from each sequencing center in the form of .bam files mapped to the 1000 Genomes hs37d5 build 37 or GRCh38DH build 38 human genome reference sequence. File transfer was via Aspera or Globus Connect, depending on the center. Batches of 100 - 500 .bam files in a single directory are convenient, along with a file of md5 checksums for the data files in that directory. The IRC validated the md5 checksum, indexed each .bam file using 'samtools index' and ran local programs Qplot (Li, et al, 2013, doi:10.1155/2013/865181) and verifyBamId (Jun, et al, 2012, doi:10.1016/j.ajhg.2012.09.004) for incoming sequence quality control. If needed, we added ''NWD'' DNA sample identifiers to the read group header lines (Illumina) and converted from UCSC to Ensembl chromosome names (Illumina and Macrogen) using 'samtools reheader'. In-house scripts were used to read group tags as needed to legacy Illumina sequencing data from 2012-2013.

To produce the harmonized read mappings which were used for variant discovery and genotyping, we remapped the sequence data in each .bam file to the 1000 Genomes GRCh38DH human genome reference sequence using the TOPMed / CCDG pipeline standard protocol. We used 'bamUtils bam2fastq' with flags '--splitRG --merge --gzip' to extract all sequence reads by read group into interleaved .fastq format and remapped to GRCh38DH using bwa mem version 0.7.15 with '-K 100000000 -Y' to produce deterministic behavior and to preserve the full sequence reads in supplementary alignments.. Samblaster v.0.1.24 added the mate MC and MQ tags. Read group header information was copied verbatim from the incoming sequencing center alignment file. This was followed by 'samtools sort', 'samtools merge' and 'bamUtils dedup_LowMem --recab --binCustom --binQualS 0:2,3:3,4:4,5:5,6:6,7:10,13:20,23:30 --allReadNames' to recalibrate and bin base call quality scores. Samtools version 1.3.1 was used throughout. Processing was coordinated and managed by in-house scripts.

Descriptions of our local and standard software tools are available from:

http://genome.sph.umich.edu/wiki/BamUtils [22]

http://genome.sph.umich.edu/wiki/GotCloud [23]

http://www.htslib.org [24] (samtools)

https://github.com/lh3/bwa [25] (bwa, current)

Software sources:

https://github.com/statgen/bamUtil/releases/tags/v1.0.1 [26] 4 [26]

https://github.com/statgen/qplot [27]

https://github.com/ [28] Griffan [28] /verifyBamI [28] D [28] /releases/tags/ [28] 1.0.1 [28]

https://github.com/samtools/samtools/ [29] archive [29] /1. [29] 3.1.zip [29]

https://github.com/lh3/bwa [25] (source code) https://github.com/lh3/bwa/tree/master/bwakit [30]

GRCh38 human genome reference source:

ftp:// ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/ [31]
GRCh38_reference_genome/GRCh38_full_analysis_set_plus_decoy_hla.fa [31]

The two sequence quality criteria we used in order to pass sequence data on for joint variant discovery and genotyping are: estimated DNA sample contamination below 3%, and fraction of the genome covered at least 10x 95% or above. DNA sample contamination was estimated from the sequencing center read mapping using an updated version of the software verifyBamId (Goo Jun, et al., 2012 Detecting and estimating contamination of human DNA samples in sequencing and array based genotype data. American Journal of Human Genetics, v.91, n.5, pp.839-848).

New procedures to access the individual level sequence data files mapped to build 38 are currently under technical development. We anticipate that controlled access for users with an approved dbGaP data access request might be provided through the dbGaP 'Run Selector' with a workflow very similar to that for the TOPMed phase 1 sequence data currently available in the NCBI Sequence Read Archive (SRA). An implementation timeline is not currently available.

Variant Discovery and Genotype Calling

Overview

Freeze 5b genotype call sets were produced by a variant calling pipeline (Figure 2) performed by the TOPMed Informatics Research Center (Center for Statistical Genetics, University of Michigan, Hyun Min Kang, Tom Blackwell and Gonçalo Abecasis). The software tools used in this version of the pipeline are available in the following repository: [32] [32] https://github.com/statgen/topmed_freeze [33] 5 [33] _calling [33] . The following description refers to specific components of the pipeline. The variant calling software tools are under continuous development; updated versions can be accessed at http://github.com/atks/vt [34] or http://github.com/hyunminkang/apigenome [35].

Figure 2 : Outline of TOPMed Freeze 5b Variant Calling Pipeline

Outline of the variant calling procedure

The GotCloud pipeline detects variant sites and calls genotypes from a list of aligned sequence reads. Specifically, the pipeline for freeze 5b consisted of the following six key steps (see also Figure 2). Most of these procedures will be integrated into the next official release of GotCloud software package.

1. **Variant detection** : For each sequenced genome (in BAM/CRAMs), candidate variants were detected by vt discover2 software tools, separated by each chromosome. The candidate variants were normalized by vt normalize algorithm.

2. **Estimation of contamination, genetic ancestry, and sex** : For each sequenced genome, genetic ancestry and DNA sequence contamination were estimated by the cramore cram-verify-bam software tool. In addition, the biological sex of each sequenced genome was inferred by computing relative depth at X

and Y chromosomes compared to the autosomal chromosomes, using the software tool cramore vcf-normalized-depth.

3. **Variant consolidation** : For each chromosome, the called variant sites were merged across the genomes, accounting for overlap of variants between genomes, using the cramore merge-candidate-variants, vt annotate_indels, vt consolidate software tool.

4. **Genotype and feature collection** : For each batch of 1,000 samples and 10Mb of chunks,the genotyping module implemented in cramore dense-genotype collects individual genotype likelihoods and variant features across the merged sites by iterating over sequenced genomes, focusing on the selected region, using the contamination levels and sex inferred in step 2. These per-batch genotypes are merged across all batches for each 100 kb region, using the cramore paste-vcf-calls software tool, producing merged and unfiltered genotypes. The estimated genetic ancestry of each individual was used as input when merging genotypes to compute variant features involving individual-specific allele frequencies.

5. Inference of nuclear pedigree : Genotypes at ~600,000 SNPs polymorphic in the Human Genome Diversity Project (HGDP) data were extracted using cramore vcf-squeeze and cramore vcf-extract tools. These genotypes and inferred sex at step 2 were used together to infer a pedigree consisting of duplicated individuals and nuclear families using king2 and vcf-infer-ped software tools.

6. **Variant filtering** : We use the inferred pedigree of related and duplicated samples to calculate Mendelian consistency statistics using vt milk-filter, and to train a variant classifier using a Support Vector Machine (SVM) implemented in the libsvm software package.

Steps to prepare input files, install software and perform variant calling

To produce variant calls using this pipeline, the following input files need to be prepared:

1. Aligned sequence reads in BAM or CRAM format. Each BAM and CRAM file should contain one sample per subject.

2. A sequence index file. Each line should contain [Sample ID] [Full Path to the BAM/CRAM file]. See data/samples.index for example.

To clone and build the repository, follow these steps

```
$ git clone https://github.com/statgen/topmed_freeze5_calling.git [36]
```

```
$ cd topmed_freeze5_calling
```

```
$ make # or make -j [numjobs] to expedite the process
```

```
$ wget ftp://anonymous@share.sph.umich.edu/gotcloud/ref/hs38DH-db142-v1.tgz [37] # this will take a while
```

```
$ tar xzvf hs38DH-db142-v1.tgz
```

```
$ rm hs38DH-db142-v1.tgz
```

After these steps, modify scripts/gcconfig.pm to specify input data files or other parameters. Modifying the first section (index and ped file in particular) should be minimally required changes.

To perform variant discovery, run the steps as documented in the GitHub repository.

```
$ perl scripts/run-discover-variantsde
```

After this step, follow these instruction to run make -f [Makefile] -j [numjobs] to complete the discovery tasks

To estimate genetic ancestries and contamination levels, run the following steps:

```
$ perl scripts/run-verify-bam-xy-depth
```

After this step, follow the instruction to run make -f [Makefile] -j [numjobs] to complete the tasks to estimate key per-sample parameters. Based on the results from this, a sex map file and a new index file that contains contamination estimates and genetic ancestry will need to be created.

To perform variant consolidation, run the following steps:

```
$ perl scripts/run-union-variants
```

After this step, follow the instruction to run make -f [Makefile] -j [numjobs] to complete the tasks to obtain the merged list of variants.

To perform genotyping by a batch of 1,000 samples for each 10Mb region, run the following steps.

```
$ perl scripts/run-batch-genotypes [whitespace separated chromosome names to call]
```

After this step, follow the instruction to run make -f [Makefile] -j [numjobs] to complete the tasks to obtain per-batch unfiltered genotypes.

To merge the per-batch genotype across all batches, run the following steps:

```
$ perl scripts/run-paste-genotypes [whitespace separated chromosome names to call]
```

After this step, follow the instruction to run make -f [Makefile] -j [numjobs] to complete the tasks to obtain merged unfiltered genotypes. This step will also produce a subset of genotypes overlapping HGDP-polymorphic variants to extract the genotypes at known good variants.

To infer a pedigree file that contains duplicates and nuclear families, run the following steps:

```
$ perl scripts/infer-ped
```

This step will also produce a PED format file that annotates duplicates samples and nuclear families.

To compute per-variant Mendelian error estimates, run the following steps:

```
$ perl scripts/run-milk [whitespace separated chromosome names to call]
```

After this step, follow the instruction to run make -f [Makefile] -j [numjobs] to complete the tasks to obtain Mendelian error estimate.

To perform variant filtering and annotation, run the following steps:

```
$ perl scripts/run-svm-filter [whitespace separated chromosome names to call]
```

After this step, follow the instruction to run make -f [Makefile] -j [numjobs] to complete the tasks to obtain a list of filtered variants.

To finalize all the steps to produce final VCF file, run the following steps:

```
$ perl scripts/produce-final-vcfs [whitespace separated chromosome names to call]
```

After this step, follow the instruction to run make -f [Makefile] -j [numjobs] to complete the tasks to obtain a list of filtered variants.

Variant Detection

Variant detection from each sequenced (and aligned) genome was performed by the vt discover2 software tool.

The variant detection algorithm considers a potential candidate variant if there exists a mismatch between the aligned sequence reads and the reference genome. Because such a mismatch can easily occur by random errors, only potential candidate variants passing the following criteria are considered to be ***candidate variants*** in the next steps.

1. At least two identical evidences of variants must be observed from aligned sequence reads.

a. Each individual evidence will be normalized using the normalization algorithm implemented in vt normalize software tools.

b. Only evidence from reads with mapping quality 20 or greater will be considered.

c. Duplicate reads, QC-failed reads, supplementary reads, and secondary reads will be ignored.

d. Evidence of a variant within overlapping fragments of read pairs will not be double counted. Either end of the overlapping read pair will be soft-clipped using bam clipOverlap software tool.

2. Assuming per-sample heterozygosity of 0.1%, the posterior probability of having a variant at the position should be greater than 50%. This method is equivalent to the glfSingle model described in [38] http://www.ncbi.nlm.nih.gov/pubmed/25884587 [38]

The variant detection step is required only once per sequenced genome, when multiple freezes of variant calls are produced over the course of time.

Variant Consolidation

Variants detected from the discovery step were merged across all samples.

1. The non-reference alleles normalized by vt normalize algorithm were merged across the samples, and unique alleles were printed as biallelic candidate variants. The algorithm is published at [39] http://www.ncbi.nlm.nih.gov/pubmed/25701572. [39]

2. For alleles overlapping with other SNPs and Indels, overlap_snp and overlap_indel filters were added in the FILTER column of the corresponding variant.

3. If there were tandem repeats with 2 or more repeats with total repeat length of 6bp or longer, the variant was annotated as a potential VNTR (Variant Number Tandem Repeat), and overlap_vntr filters were added to the variant overlapping with the repeat track of the putative VNTR.

Variant Genotyping and Feature Collection

The genotyping step iterates all of the merged variant sites over the sequenced samples. It iterates over BAM/CRAM files one at a time sequentially for each 1Mb chunk to perform contamination-adjusted genotyping and annotation of variant features for filtering. The following variant features are calculated during the genotyping procedure.

- AVGDP : Average read depth per sample
- AC : Non-reference allele count
- AN : Total number of alleles
- GC : Genotype count
- GN : Total genotype counts
- HWE_AF : Allele frequency estimated from PL under HWE
- FIBC_P : [Obs(Het) - Exp(Het)] / Exp[Het] without correcting for population structure
- FIBC_I : [Obs(Het) - Exp(Het)] / Exp[Het] after correcting for population structure
- HWE_SLP_P : -log(HWE score test p-value without correcting for population structure) □ sign(FIBC_P)
- HWE_SLP_I : -log(HWE score test p-value after correcting for population structure) □ sign(IS_IBC)
- MIN_IF : Minimum value of individual-specific allele frequency
- MAX_IF : Maximum value of individual-specific allele frequency
- ABE : Average fraction [#Ref Allele] across all heterozygotes
- ABZ : Z-score for testing deviation of ABE from expected value (0.5)
- BQZ: Z-score testing association between allele and base qualities
- CYZ: Z-score testing association between allele and the sequencing cycle
- STZ : Z-score testing association between allele and strand
- NMZ : Z-score testing association between allele and per-read mismatches
- IOR : log [Obs(non-ref, non-alt alleles) / Exp(non-ref, non-alt alleles)]
- NM1 : Average per-read mismatches for non-reference alleles
- NM0 : Average per-read mismatches for reference alleles

The genotyping process was done with adjustment for potential contamination. It uses adjusted genotype likelihood similar to the published method [40] https://github.com/hyunminkang/cleancall [40] , but does not use estimated population allele frequency for the sake of computational efficiency. It conservatively assumes that the probability of observing a non-reference read given a homozygous reference genotype is equal to half of the estimated contamination level, (or 1%, whichever is greater). The probability of observing a reference read given a homozygous non-reference genotype was calculated in a similar way. This adjustment makes the heterozygous call more conservatively when the reference and non-reference allele reads are strongly imbalanced. For example, if 45 reference alleles and 5 non-reference alleles are observed at Q40, the new method calls it as homozygous reference genotype while the original method ignoring potential contamination calls it as heterozygous genotype. This adjustment improves the genotype quality of contaminated samples by reducing genotype errors by several fold.

Variant Filtering

The variant filtering in TOPMed Freeze 4 were performed by (1) first calculating Mendelian consistency scores using known familial relatedness and duplicates, and (2) training a Support Vector Machine (SVM) classifier between the known variant sites (positive labels) and the Mendelian inconsistent variants (negative labels).

Negative labels were defined if the Bayes Factor for Mendelian consistency quantified as Pr(Reads | HWE, Pedigree) / Pr(Reads | HWD, no Pedigree) is less than 0.001. Also a variant was marked as negative labels if 3 or more samples showed 5% of non-reference Mendelian discordance within families or genotype discordance between duplicated samples. The positive labels were the SNPs found polymorphic either in the 1000 Genomes Omni2.5 array or in HapMap 3.3, with additional evidence of being polymorphic from the sequenced samples. Variants eligible to be marked with both positive and negative labels were discarded from the labels. The SVM scores trained and predicted by the libSVM software tool were annotated in the VCF file.

Two additional hard filters were applied. (1) Excess heterozygosity filter (EXHET), if the Hardy-Weinberg disequilibrium p-value was less than 1e-6 in the direction of excess heterozygosity after accounting for population structure. An additional ~3,900 variants were filtered out by this filter. (2) Mendelian discordance filter (DISC), with 3 or more Mendelian inconsistencies or duplicate discordances observed from the samples. An additional ~370,000 variants were filtered out by this filter.

Functional annotation for each variant is provided in the INFO field using Pablo Cingolani's snpEff 4.1 with a GRCh38.76 database. The current release includes only hard-call genotypes in the VCF files, without genotype likelihoods and with no missing genotypes. Genotype likelihoods may be included in future releases, at the cost of approximately 100x greater file size.

Data Coordinating Center Methods

Cathy Laurie, Bruce Weir and Ken Rice

Genetic Analysis Center, Department of Biostatistics, University of Washington

The following three approaches were used to identify and resolve sample identity issues.

Concordance between annotated sex and genetic sex inferred from the WGS data

Genetic sex was inferred from normalized X and Y chromosome depth for each sample (i.e. divided by autosomal depth) and from X chromosome heterozygosity. A small number of sex mismatches were detected as annotated females with low X and high Y chromosome depth or annotated males with high X and low Y chromosome depth. These samples were either excluded from the sample set to be released on dbGaP or their sample identities were resolved using information from prior array genotype comparisons or pedigree checks. We also identified a small number of apparent sex chromosome aneuploidies (XXY, XXX, XYY and mosaics such as XX/XO and XY/XO). If applicable to a study, these are annotated in a file accompanying the genotypes, with the suffix "sex_chromo_karyotype.txt". This file provides the NWD ID, genetic sex, and inferred sex chromosome karyotype of samples with apparent aneuploidies.

Concordance between prior SNP array genotypes and WGS-derived genotypes

Prior genome-wide SNP array data are available for 27 of the 32 accessions to be released in 2018 (all except phs001143, phs001024, phs001062, phs001032, and phs000997). The average percentage of individuals within the 27 accessions that have prior array data is 85%.

For 10 accessions, the prior array data analyzed for TOPMed were derived from 'fingerprints' compiled by dbGaP (Yumi Jin, see URL below); these fingerprints consist of genotypes from a set of 10,000 bi-allelic autosomal SNP markers chosen to occur on multiple commercial arrays and to have a minor allele frequency (MAF) > 5%. For another 16 accessions, all autosomal SNPs with MAF > 5% on a genome-wide array were used. Two studies had a combination of fingerprint and array data. For either fingerprint and/or full array data, percent concordance with WGS was determined by matching on heterozygous versus homozygous status (rather than specific alleles) to avoid strand issues. Concordance percentages for array-WGS matches were generally in the high 90s, while those considered to be mismatches were in the 50-60% range (empirically determined to be the expected matching level for random pairs of samples). We found that >99% of the WGS samples tested were concordant with prior array data. Discordant samples were either excluded from the 2018 release or resolved as sample switches using pedigree and/or sex-mismatch results.

SNP fingerprints: [41] [41] http://www.ashg.org/2014meeting/abstracts/fulltext/f140122979.htm [41]

Comparisons of observed and expected relatedness from pedigrees

Kinship coefficients (KCs) were estimated for all pairs of individuals using ~250k single nucleotide variants that are autosomal, MAF >5%, and pruned to have low linkage disequilibrium (r^2<0.1) with one another. The estimation procedure used 'PC-Relate' (Conomos et al. 2016, DOI: [42] 10.1016/j.ajhg.2015.11.022 [42]), which is robust to population structure, admixture and departures from Hardy-Weinberg Equilibrium. The KC estimates were compared to those expected from pedigrees for the accessions with annotated family structure (phs000956, phs000974, phs000988, phs000964, phs000954 , phs001143, phs001207, phs001215, phs001217, phs001218, phs001293, phs001345, phs001359, phs001387, phs001412, and phs001416). Discrepancies between observed and expected KCs were investigated and, in many cases, resolved either by correcting sample-subject mapping for sample switches or by revising the pedigree structure. Pedigree changes were warranted when one alteration resolved multiple KC discrepancies or when supported by additional information from the studies.

Source URL (modified on 10/28/2021 - 10:14am):https://topmed.nhlbi.nih.gov/topmed-whole-genome-sequencing-project-freeze-5b-phases-1-and-2
Links
[1] https://topmed.nhlbi.nih.gov/topmed-whole-genome-sequencing-project-freeze-5b-phases-1-and-2#overview [2] https://topmed.nhlbi.nih.gov/topmed-whole-genome-sequencing-project-freeze-5b-phases-1-and-2#Summary5b [3] https://topmed.nhlbi.nih.gov/topmed-whole-genome-sequencing-project-freeze-5b-phases-1-and-2#ids [4] https://topmed.nhlbi.nih.gov/topmed-whole-genome-sequencing-project-freeze-5b-phases-1-and-2#controlsamples [5]

https://topmed.nhlbi.nih.gov/topmed-whole-genome-sequencing-project-freeze-5b-phases-1-and-2#seqctrmethods [6] https://topmed.nhlbi.nih.gov/topmed-whole-genome-sequencing-project-freeze-5b-phases-1-and-2#broad [7] https://topmed.nhlbi.nih.gov/topmed-whole-genome-sequencing-project-freeze-5b-phases-1-and-2#nwgc [8] https://topmed.nhlbi.nih.gov/topmed-whole-genome-sequencing-project-freeze-5b-phases-1-and-2#nygc [9] https://topmed.nhlbi.nih.gov/topmed-whole-genome-sequencing-project-freeze-5b-phases-1-and-2#illumina [10] https://topmed.nhlbi.nih.gov/topmed-whole-genome-sequencing-project-freeze-5b-phases-1-and-2#macrogen [11] https://topmed.nhlbi.nih.gov/topmed-whole-genome-sequencing-project-freeze-5b-phases-1-and-2#bcmhgsc [12] https://topmed.nhlbi.nih.gov/topmed-whole-genome-sequencing-project-freeze-5b-phases-1-and-2#ircmethods [13] https://topmed.nhlbi.nih.gov/topmed-whole-genome-sequencing-project-freeze-5b-phases-1-and-2#readalignments [14] https://topmed.nhlbi.nih.gov/topmed-whole-genome-sequencing-project-freeze-5b-phases-1-and-2#variantdiscoveryandgtcalling [15] https://topmed.nhlbi.nih.gov/topmed-whole-genome-sequencing-project-freeze-5b-phases-1-and-2#dccmethods [16] https://topmed.nhlbi.nih.gov/topmed-whole-genome-sequencing-project-freeze-5b-phases-1-and-2#concordancesex [17] https://topmed.nhlbi.nih.gov/topmed-whole-genome-sequencing-project-freeze-5b-phases-1-and-2#concordancegenotypes [18] https://topmed.nhlbi.nih.gov/topmed-whole-genome-sequencing-project-freeze-5b-phases-1-and-2#comparisons [19] http://topmed.nhlbi.nih.gov/ [20] https://topmed.nhlbi.nih.gov/ [21] https://github.com/Illumina [22] http://genome.sph.umich.edu/wiki/BamUtils [23] http://genome.sph.umich.edu/wiki/GotCloud [24] http://www.htslib.org/ [25] https://github.com/lh3/bwa [26] https://github.com/statgen/bamUtil/releases/tags/v1.0.14 [27] https://github.com/statgen/qplot [28] https://github.com/Griffan/verifyBamID/releases/tags/1.0.1 [29] https://github.com/samtools/samtools/archive/1.3.1.zip [30] https://github.com/lh3/bwa/tree/master/bwakit [31] http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.fa.gz [32] https://github.com/statgen/topmed_freeze3_calling [33] https://github.com/statgen/topmed_freeze5_calling [34] http://github.com/atks/vt [35] http://github.com/hyunminkang/apigenome [36] https://github.com/statgen/topmed_freeze5_calling.git [37] ftp://anonymous@share.sph.umich.edu/gotcloud/ref/hs38DH-db142-v1.tgz [38] http://www.ncbi.nlm.nih.gov/pubmed/25884587 [39] http://www.ncbi.nlm.nih.gov/pubmed/25701572 [40] https://github.com/hyunminkang/cleancall [41] http://www.ashg.org/2014meeting/abstracts/fulltext/f140122979.htm [42] https://dx.doi.org/10.1016/j.ajhg.2015.11.022