TOPMed Whole Genome Sequencing Methods: Freeze 8

August 2019

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 >80 different studies with ~158,000 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 (www.nhlbiwgs.org).

Studies have been added to the TOPMed program in approximately five yearly "Phases", with Phase 1 beginning in October 2014, and Phase 5 in October 2018. WGS data are acquired and reads aligned to the reference genome more or less continuously by multiple Sequencing Centers. Periodically, the TOPMed Informatics Research Center (IRC) performs variant identification and genotype calling on all samples available at a given time and the resulting call set is referred to as a genotype "Freeze".

Some studies have samples sequenced through both the NHLBI TOPMed program and the NHGRI Centers for Common Disease Genetics (CCDG) program. For these studies, the joint variant identification and genotype calling performed by the IRC includes both TOPMed- and CCDG-funded samples sequenced within the same time frame.

This document contains descriptions of Sequencing Center methods for TOPMed Phases 1-4 and CCDG samples for studies in common between the two programs. It also contains descriptions of joint variant identification and genotype calling performed by the IRC in Freeze 8. 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 within a given Phase were sequenced at the same center. The reads were aligned to human genome build GRCh38 using a common pipeline across all centers. The IRC performed joint genotype calling on all samples in Freeze 8. 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 by dbGaP in VCF format, with one file per chromosome. GRCh38 read alignments (as CRAM files) are stored in a cloud environment to which access is provided through a virtual directory service managed by NCBI’s Sequence Read Archive (with data access permissions inherited from TOPMed dbGaP accessions). See section “Access to sequence data”.

A summary of the dbGaP accessions for studies included in Freeze 8, including their approximate sample numbers and Sequencing Center, 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. The TOPMed web site (https://www.nhlbiwgs.org/topmed-data-access-scientific-community) provides further information about data structures and access.

Table 1. Summary of TOPMed dbGaP Study Accessions in Freeze 8

TOPMed Project[1]	dbGaP TOPMed Study Accession	Study Name[2]	Study Abbreviation	Study PI	Sample Size[3]	Sequencing Center[4]	TOPMed Phase	dbGaP Parent Study Accession
AFGen	phs001543	Atrial Fibrillation Biobank Ludwig Maximilian University Study	AFLMU	Sinner, Mortiz; Kaab, Stefan	350	BROAD	2.5, TOPMed/ CCDG	
Amish	phs000956	Genetics of Cardiometabolic Health in the Amish	Amish	Mitchell, Braxton D	1,118	BROAD	1	
AFGen, VTE	phs001211	Atherosclerosis Risk in Communities Study VTE cohort	ARIC	Boerwinkle, Eric	8,452	BAYLOR, BROAD	1, 2, CCDG	phs000280
AFGen	phs001435	Molecular Mechanisms of Inherited Cardiomyopathies and Arrhythmias in the Australian	AustralianFamilialAF	Fatkin, Diane	120	BROAD	1.5	

		Familial AF Study						
BAGS	phs001143	New Approaches for Empowering Studies of Asthma in Populations of African Descent - Barbados Asthma Genetics Study	BAGS	Barnes, Kathleen	1,086	ILLUMINA	1	
BioMe	phs001644	Mount Sinai BioMe Biobank	BioMe	Loos, Ruth; Kenny, Eimear	11,597	BAYLOR, WASHU	3, TOPMed/ CCDG, CCDG	phs000925
AFGen	phs001624	The Vanderbilt University BioVU Atrial Fibrillation Genetics Study	BioVU_AF	Shoemaker, M. Benjamin; Roden, Dan	1,129	BAYLOR	TOPMed/ CCDG	
CRA_CAMP	phs001726	Childhood Asthma	CAMP	Weiss, Scott	1,785	UW	3	phs000166

		Management Program						
CARDIA	phs001612	Coronary Artery Risk Development in Young Adults	CARDIA	Fornage, Myriam; Hou, Lifang;Lloyd-Jones, Donald	3,456	BAYLOR	3	phs000285
ATGC	phs001728	Childhood Asthma Research and Education Network: Best Add-on Therapy Giving Effective Responses	CARE_BADGER	Martinez, Fernando	36	UW	3	
ATGC	phs001729	Childhood Asthma Research and Education Network: Characterizing the Response to a Leuikotriene Receptor Agonis and an Inhaled Corticosteroid	CARE_CLIC	Martinez, Fernando	13	UW	3	phs000166

ATGC	phs001730	Childhood Asthma Research and Education Network: Pediatric Asthma Controller Trial	CARE_PACT	Martinez, Fernando	22	UW	3	phs000166
ATGC	phs001732	Childhood Asthma Research and Education Network: Treating Children to Prevent Exacerbations of Asthma	CARE_TREXA	Martinez, Fernando	67	UW	3	
AFGen	phs001600	The Duke CATHeterization GENetics Study	CATHGEN	Kraus , William; Sun, Albert	123	BROAD	TOPMed/ CCDG	phs000703
AFGen	phs001189	Cleveland Clinic Atrial Fibrillation Study	CCAF	Chung, Mina; Barnard, John	364	BROAD	1	phs000820

CFS	phs000954	Cleveland Family Study - WGS Collaboration	CFS	Redline, Susan	1,296	UW	1, 3.5	phs000284
ATGC	phs001602	Children's Health Study: Integrative Genetic Approaches to Gene-Air Pollution Interactions in Asthma	ChildrensHS_ GAP	Gilliland, Frank	7	UW	3	
ATGC	phs001603	Children's Health Study: Integrative Genomics and Environmental Research of Asthma	ChildrensHS_ IGERA	Gilliland, Frank	157	UW	3	
ATGC	phs001604	Children's Health Study: Effects of Air Pollution on the Development of Obesity in Children	ChildrensHS_ MetaAir	Gilliland, Frank	52	UW	3	

ATGC	phs001605	Genetics Sub-Study of Chicago Initiative to Raise Asthma Health Equity	CHIRAH	Kumar, Rajesh	269	UW	3	
CHS, VTE	phs001368	Cardiovascular Health Study	CHS	Psaty, Bruce; Tracy, Russell	3,540	BAYLOR	2, 3	phs000287
COPD	phs000951	Genetic Epidemiology of COPD Study	COPDGene	Silverman, Edwin	10,605	BROAD, UW	1, 2, 2.5	phs000179
CRA_CAMP	phs000988	The Genetic Epidemiology of Asthma in Costa Rica - Asthma in Costa Rica cohort	CRA	Weiss, Scott	3,217	UW	1, 3	
AFGen	phs001546	Determining the association of chromosomal variants with non-PV triggers and ablation-outcome in DECAF	DECAF	Mohanty, Sanghamitra; Natale, Andrea	6	BROAD	1.5	

AA_CAC	phs001412	Diabetes Heart Study	DHS	Palmer, Nicholette; Bowden, Donald W.	394	BROAD	2	phs001012
ECLIPSE	phs001472	Evaluation of COPD Longitudinally to Identify Predictive Surrogate End-points	ECLIPSE	Silverman, Edwin	2,345	WASHU	3	phs001252
AFGen	phs001606	Estonian Genome Center	EGCUT	Esko, Tonu	324	BROAD	2.5	
COPD	phs000946	Boston Early-Onset COPD Study	EOCOPD	Silverman, Edwin	145	BROAD, UW, WASHU	1	phs001161
AFGen, FHS	phs000974	Framingham Heart Study	FHS	Vasan, Ramachandran S.; Cupples, L. Adrienne	4,185	BROAD	1	phs000007
ATGC	phs001542	ATGC Gene-Environment, Admixture and Latino	GALAI	Burchard, Esteban	948	UW	3	

		Asthmatics Study I Asthma						
ATGC, PGX_Asthma	phs000920	Gene-Environment, Admixture and Latino Asthmatics Study	GALAII	Burchard, Esteban	4,944	ILLUMINA[5], NYGC, UW	1, 3, CCDG, legacy[5]	phs001180
ATGC	phs001661	Genetics of Complex Pediatric Disorders - Asthma	GCPD-A	Hakonarson, Hakon	4,441	UW	3	
AFGen	phs001547	The GENetics in Atrial Fibrillation Study	GENAF	Christophersen, Ingrid Elisabeth	90	BROAD	TOPMed/ CCDG	
AA_CAC, GeneSTAR	phs001218	Genetic Studies of Atherosclerosis Risk	GeneSTAR	Mathias, Rasika	1,766	BROAD, ILLUMINA[5], MACROGEN	2, legacy[5]	phs001074

AA_CAC, HyperGEN_GENOA	phs001345	Genetic Epidemiology Network of Arteriopathy	GENOA	Kardia, Sharon; Smith, Jennifer	1,254	BROAD, UW	2	phs001238
GenSalt	phs001217	Genetic Epidemiology Network of Salt Sensitivity	GenSalt	He, Jiang	1,858	BAYLOR	2	phs000784
AFGen	phs001725	Groningen Genetics of Atrial Fibrillation Study	GGAF	Rienstra, Michiel	640	BAYLOR	TOPMed/ CCDG	
GOLDN	phs001359	Genetics of Lipid Lowering Drugs and Diet Network	GOLDN	Arnett, Donna K	965	UW	2	phs000741
HCHS_SOL	phs001395	Hispanic Community Health Study - Study of Latinos	HCHS_SOL	Kaplan, Robert; North, Kari	7,963	BAYLOR	3, CCDG	phs000810
AFGen, VTE	phs000993	Heart and Vascular Health Study	HVH	Heckbert, Susan	699	BAYLOR, BROAD	1, 2	phs001013

HyperGEN_GENOA	phs001293	Hypertension Genetic Epidemiology Network	HyperGEN	Arnett, Donna K	1,898	UW	2	
AFGen	phs001545	Intermountain Heart Study	INSPIRE_AF	Cutler, Michael; Johnson, Kevin; Schwab, Angie	466	BROAD	TOPMed/ CCDG	
IPF	phs001607	Whole Genome Sequencing in Familial and Sporadic Idiopathic Pulmonary Fibrosis	IPF	Schwartz, David; Fingerlin, Tasha	1,495	WASHU	3	
JHS	phs000964	Jackson Heart Study	JHS	Correa, Adolfo; Wilson, James	3,418	UW	1	phs000286
AFGen	phs001598	The Johns Hopkins University School of Medicine Atrial Fibrillation Genetics Study	JHU_AF	Nazarian, Saman	290	BROAD	TOPMed/ CCDG	

LTRC	phs001662	Lung Tissue Research Consortium	LTRC	Barwick, Lucas	1,413	BROAD	4	
VTE	phs001402	Mayo Clinic Venous Thromboembolism Study	Mayo_VTE	de Andrade, Mariza	1,350	BAYLOR	2	phs000289
AA_CAC, MESA	phs001416	Multi-Ethnic Study of Atherosclerosis	MESA	Rotter, Jerome; Rich, Stephen	5,382	BROAD	2	phs000209
AFGen	phs001062	Massachusetts General Hospital Atrial Fibrillation Study	MGH_AF	Ellinor, Patrick	1,104	BROAD	1, 1.4, 1.5, 1.6, TOPMed/ CCDG	phs001001
AFGen	phs001434	Defining time-dependent genetic and transcriptomic responses to cardiac injury among patients with arrhythmias	miRhythm	McManus, David	65	BROAD	1.5	

MLOF	phs001515	My Life, Our Future: Genotyping for Progress in Hemophilia	MLOF	Konkle, Barbara; Johnsen, Jill	5,137	BAYLOR, NYGC	2, 3	
AFGen	phs001544	Malmo Preventative Project	MPP	Smith, Gustav	121	BROAD	TOPMed/ CCDG	
OMG_SCD	phs001608	Outcome Modifying Genes in Sickle Cell Disease	OMG_SCD	Ashley-Koch, Allison; Telen, Marilyn	653	BAYLOR	2	
AFGen	phs001024	Partners Healthcare Biorepository	Partners	Lubitz, Steven	128	BROAD	1	
PCGC_CHD	phs001735	Pediatric Cardiac Genomics Consortium's Congenital Heart Disease Biobank	PCGC_CHD	Gelb, Bruce; Seidman, Christine	1,906	BROAD	4	phs001194
PharmHU	phs001466	The Pharmacogenomics of	PharmHU	Boerwinkle, Eric; Sheehan, Vivien	862	BAYLOR	2	

		Hydroxyurea in Sickle Cell Disease						
ATGC	phs001727	Pathways to Immunologically Mediated Asthma	PIMA	Martinez, Fernando	56	UW	3	
AFGen	phs001601	Early-onset Atrial Fibrillation in the Penn Medicine BioBank Cohort	PMBB_AF	Rader, Daniel; Damrauer, Scott	421	BROAD	TOPMed/ CCDG	
PUSH_SCD	phs001682	Pulmonary Hypertension and the Hypoxic Response in Sickle Cell Disease	PUSH_SCD	Nekhai, Sergei	423	BAYLOR	2	
REDS-III_Brazil	phs001468	Recipient Epidemiology and Donor Evaluation Study-III	REDS-III_Brazil	Custer, Brian; Kelly, Shannon	2,634	BAYLOR	2	

SAFS	phs001215	Whole Genome Sequencing to Identify Causal Genetic Variants Influencing CVD Risk - San Antonio Family Studies	SAFS	Blangero, John; Curran, Joanne	1,818	ILLUMINA[5]	1, legacy[5]	phs000847
ATGC, PGX_Asthma	phs000921	Study of African Americans, Asthma, Genes and Environment	SAGE	Burchard, Esteban	2,032	ILLUMINA[5], NYGC, UW	1, 3, legacy[5]	
ATGC	phs001467	Study of Asthma Phenotypes & Pharmacogenomic Interactions by Race-Ethnicity	SAPPHIRE_asthma	Williams, L. Keoki	4,793	UW	3	
Sarcoidosis	phs001207	Genetics of Sarcoidosis in African Americans	Sarcoidosis	Montgomery, Courtney	933	BAYLOR, UW	2, 3.5	

SARP	phs001446	Severe Asthma Research Program	SARP	Meyers, Deborah A	1,890	NYGC	2	phs000422
SAS	phs000972	Samoan Adiposity Study	SAS	McGarvey, Stephen	1,294	NYGC, UW	1, 2	phs000914
THRV	phs001387	Taiwan Study of Hypertension using Rare Variants	THRV	Rao, D.C.; Chen, Yii-Der Ida	2,170	BAYLOR	2	
AFGen	phs000997	Vanderbilt Atrial Fibrillation Ablation Registry	VAFAR	Shoemaker, M. Benjamin	176	BROAD	1, 1.5, TOPMed/ CCDG	
AFGen	phs001032	Vanderbilt Genetic Basis of Atrial Fibrillation	VU_AF	Darbar, Dawood	1,140	BROAD	1	phs000439
walk_PHaSST	phs001514	Treatment of Pulmonary Hypertension and Sickle Cell Disease with Sildenafil Therapy	walk_PHaSST	Gladwin, Mark;Zhang, Yingze	437	BAYLOR	2	

AFGen	phs001040	Women's Genome Health Study	WGHS	Albert, Christine; Chasman, Daniel	118	BROAD	1	
WHI	phs001237	Women's Health Initiative	WHI	Kooperberg, Charles; Reiner, Alex	11,071	BROAD	2	phs000200

1 - AA_CAC=African American Coronary Artery Calcification project; AFGen=Identification of Common Genetic Variants for Atrial Fibrillation and PR Interval - Atrial Fibrillation Genetics Consortium; Amish=Genetics of Cardiometabolic Health in the Amish; ATGC=Asthma Translational Genomics Collaborative; BAGS=Barbados Asthma Genetics Study; BioMe=Mount Sinai BioMe Biobank; CARDIA=Whole Genome Sequence Analysis in Early Cerebral Small Vessel Disease; CCDG=Whole Genome Sequence Analysis in Early Cerebral Small Vessel Disease; CFS=Cleveland Family Study; CHS=Cardiovascular Health Study; COPD=Genetic Epidemiology of COPD; CRA_CAMP=The Genetic Epidemiology of Asthma in Costa Rica and the Childhood Asthma Management Program; ECLIPSE=Evaluation of COPD Longitudinally to Identify Predictive Surrogate Endpoints; FHS=Framingham Heart Study; GeneSTAR=Genetic Studies of Atherosclerosis Risk; GenSalt=Genetic Epidemiology Network of Salt Sensitivity; GOLDN=Genetics of Lipid Lowering Drugs and Diet Network; HCHS_SOL=Hispanic Community Health Study - Study of Latinos; HyperGEN_GENOA=Hypertension Genetic Epidemiology Network and Genetic Epidemiology Network of Arteriopathy; IPF=Whole Genome Sequencing in Familial and Sporadic Idiopathic Pulmonary Fibrosis; JHS=Jackson Heart Study; LTRC=Lung Tissue Research Consortium; MESA=Multi-Ethnic Study of Atherosclerosis; MLOF=My Life, Our Future: Genotyping for Progress in Hemophilia; OMG_SCD=Outcome Modifying Genes in Sickle Cell Disease; PCGC_CHD=Pediatric Cardiac Genomics Consortium's Congenital Heart Disease Biobank; PGX_Asthma=Pharmacogenomics of Bronchodilator Response in Minority Children with Asthma; PharmHU=The Pharmacogenomics of Hydroxyurea in Sickle Cell Disease; PUSH_SCD=Pulmonary Hypertension and the Hypoxic Response in Sickle Cell Disease; REDS-III_Brazil=Recipient Epidemiology and Donor Evaluation Study-III; SAFS=San Antonio Family Studies; Sarcoidosis=Genetics of Sarcoidosis in African Americans; SARP=Severe Asthma Research Program; SAS=Samoan Adiposity Study; THRV=Taiwan Study of Hypertension using Rare Variants; VTE=Venous Thromboembolism project; walk_PHaSST=Treatment of Pulmonary Hypertension and Sickle Cell Disease with Sildenafil Therapy; WHI=Women's Health Initiative. Project descriptions are available on the TOPMed website, https://www.nhlbiwgs.org.

2 - Study name as it appears in dbGaP

3 - Approximate sample size for freeze8 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; WASHU = McDonnell Genome Institute

5 - ILLUMINA was an additional sequencing center for legacy data contributed by GALAII (n=6 samples), SAGE (n=10 samples), GeneSTAR (n=283 samples), and SAFS (n=626 samples).

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 dbGaP 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. 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 in all Phases. 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 methods described below showcase the process for Phase 1 and highlight the changes that were implemented during subsequent Phases. Note that the Broad did not receive a WGS allotment for Phase 3.

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. 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, and for all Phase 4 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 (Solid Phase Reversible Immobilization) 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 (Integrated DNA Technologies)). On March 29, 2019, there was a switch from using indexed adapters provided by IDT to indexed adapters purchased directly from Illumina (product 20022370). 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 Phases 2 and 4, 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

TOPMed Phases 1, 2, and 4 followed the same process except for version changes in the software. 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 Phases 2 and 4, 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 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 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 HiSeq 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 Phases 2 and 4, 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 and continued to be used in Phase 4.

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 was analyzed on the cloud as part of Phases 2 and 4, we used the following versions for aggregation and alignment to Homo_sapiens_assembly38: picard MarkDuplicates version 2.18.15, 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 for Phases 1 and 2. For Phase 4, additional metrics were required to be met: mean coverage >=30x, 20X % >=90% and 10X % >=95%. Also, the target for PF HQ Aligned Q20 Bases was >= 8.6 x 10^10 bases. 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 a 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 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, 2, and 3. The methods given below were the same for all Phases except where noted otherwise. For Phase 1, all samples were sequenced at Macrogen (with methods described in this section); for Phase 2 and 3, 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 380 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 protocols for TruSeq PCR-free Kit (Illumina, cat# FC-121-3003) for Phase 1 studies, and by KAPA Hyper Prep Kit without amplification (KR0961.v1.14) for Phase 2 and 3 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

For Phase 1, eight normalized and indexed libraries were pooled together and denatured before cluster generation on a cBot. For subsequent phases the pool size was increased from 8-plex to 9-plex pools. The multi-plex pools were loaded on eight lanes of a flow cell and sequenced on a HiSeq X using illumina's HiSeq X reagents kit (V2.5, cat# FC-501-2521). For cluster generation, every step was controlled by the 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. Samples sequenced at Macrogen were transferred to NWGC for alignment, merging, variant calling and sequencing QC.

Read Processing

For Phases 1 and 2, 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; v2.6.0) (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. The method for read-processing was the same for Phase 3 but updated software versions were used for BWA-MEM (Burrows-Wheeler Aligner; v0.7.15) and GATK IndelRealigner and BaseRecalibrator (v3.7) and the human reference was updated to GRCh38.

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) fraction of genome covered greater than 20x for Phase 3 only; (4) duplicate rate; (5) mean insert size; (6) contamination ratio; (7) mean Q20 base coverage; (8) Transition/Transversion ratio (Ti/Tv); (9) fingerprint concordance > 99%; and (10) 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 and CCDG (see Table 1). The methods were the same for Phases 1 and 2 and CCDG samples, 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 were 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, the Kappa Hyper Library Preparation Kit (PCR-free) for Phase 2 samples, and with 1ug DNA input for the TruSeq PCR DNA HT Library Preparation Kit for CCDG samples -- 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 (or 9 samples per pool for CCDG), with each sample pool sequenced across 8 flow cell lanes. A 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 and CCDG) sequencing chemistry, and Illumina HiSeq Control Software v3.1.26 (Phase 1), HCS v3.3.39 (Phase 2), or HCS v3.3.76 (CCDG).

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 CCDG), and sequencing data was aligned to human reference build 37 (hs37d5 with decoy), and build 38 (GRCh38 with decoy) for CCDG, using BWA-MEM (v0.7.8 for Phase 1, v0.7.12 for Phase 2, and v0.7.15 for CCDG). Data was further processed using the GATK best-practices pipeline (v3.2-2 for Phase 1, v3.4-0 for Phase 2, and v3.5 for CCDG), with duplicate marking using Picard tools (v1.83 for Phase 1, v1.137 for Phase 2, and v2.4.1 for CCDG), realignment around indels (for Phases 1, 2 only), and base quality recalibration. Individual sample BAM files for Phases 1 and 2 were squeezed using Bamutil v1.0.9 with default parameters -- removing OQ's, retaining duplicate marking and binning quality scores (binMid) -- while CCDG sample BAM files were converted to CRAM using Samtools v1.3.1 with default parameters. Sample files were 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 (b37) or ~2.75Gb (b38) sex specific mappable genome, and that uniformity of coverage was acceptable (>90% of genome covered >20x). Sample identity and sequencing data quality were 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 genotype call sets.

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) 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, 2, and 3 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, 3, and 5 studies as well as CCDG studies (see Table 1), using the following methods. TOPMed phases 2 and 3 and CCDG samples were processed using the same protocols. Protocol changes implemented in TOPMed Phase 5 are highlighted in each section.

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. For TOPMed Phase 5 samples, commercially available Illumina TruSeq UD Indexes (Cat # 20022370) were used for preparing 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 240-350 pM. 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. Samples were loaded on the HiSeq X to achieve a minimum coverage of 30X, or 90 Gbp of unique reads aligning to the human reference per sample.

For phase 5 samples, TOPMed libraries were sequenced on NovaSeq 6000 instruments to generate WGS 150 bp, dual indexed and paired-end sequence reads. Libraries were pooled following a two-step process: a first calibration pool to assess pool uniformity and QC samples, and a re-pool, to achieve a minimum of 30x sequence coverage per sample. The multiplex library pools were run on the NovaSeq S4 using the Xp workflow. Typical loading concentrations range between 400-550pM. Real-time analysis (RTA) software (RTA v3.3.3) was used to monitor run performance, assessing cluster density, signal intensity and phasing/pre-phasing data.

HGSC-LIMS tracks sequence run set-up, status and the battery of performance metrics. Each of these metrics were 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 analysis and 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 GRCh38 reference genome. 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.

McDonnell Genome Institute (MGI) at Washington University

Susan Dutcher

The MGI sequenced TOPMed Phase 3 and CCDG samples (see Table 1), using the following methods.

DNA Sample Handling and QC

To ensure accuracy, project management technicians matched the nomenclature found on the sample storage vessel (tube, box, tray, etc.) to sample intake documentation provided for each shipment. In order to enter the production pipeline, the technicians associated each sample with a work order (WO) that specified in advance the path that the sample would follow (e.g., WGS, whole exome sequencing), the SOP to be followed, the desired coverage, read type, and read lengths, the post-production read alignment target, quality checking procedures, and reports to be generated. All subsequent pipeline activities involving each sample required a barcode scan recorded by our Laboratory Information Management System (LIMS). The project title and organism-naming system was information used as input in order to create barcodes for all samples. Existing barcodes on storage devices were also directly linked to the sample ID created within LIMS. Finally, technicians from the project management team send a notification to the Resource Bank. The customized LIMS is used to track the transfers of materials between production operations groups through a barcoded check in/check out process. 2D barcode sample storage tubes are utilized throughout the process.

The Resource Bank technicians assessed all DNA samples before they enter the production pipeline. Quality control measures included a re-assessment of a small number of all aliquot volumes. The library construction technicians received qualified samples for processing. A series of scripts enabled re-arraying of the TOPMed and CCDG samples based on a specified order which included case control design and the randomized order of samples. Boxes were arranged on an automated platform and individual tubes were scanned and positions were identified on the automation deck. The technicians used the specified sample lists and orientation information to allow for the re-array of the samples. Specific QC measures for TOPMed samples included quantitation using the Qubit Flourometer 3.0 with the dsDNA High Sensitivity Kit or the Varioskan Fluorometer with the Quant-iT PicoGreen dsDNA Kit. An Illumina Infinium array was used to generate genotyping data which was used for sample integrity confirmation after data generation.

Library Construction

For the TOPMed Phase 3 and CCDG projects, library core technicians used 600ng of DNA starting material with a minimum starting amount of 450ng. The library construction technicians fragmented the DNA with the Covaris LE220 Focused Ultrasonicator. A mean fragment size of ~375 bp was achieved. The library core technicians then prepared KAPA Hyper PCR-free libraries (Roche) using Perkin Elmer SciClone NGS (96-well configuration). The dual-same index system from Illumina was utilized to ensure DNA molecule integrity and data quality, which allows identification and elimination of chimeric molecules formed by cross-sample "index hopping". The library construction technicians assessed the libraries for quality and quantity using the HT DNA Hi Sens Dual Protocol Assay with the DNA Extended Range LabChip (GX, GX Touch HT) on the Perkin Elmer LabChip GX instrument using the manufacturer's instructions. The library construction team passed undiluted libraries to the loading team for qPCR, pooling, and sequencing.

Clustering and Sequencing

The HiSeq X cluster was used to generate the data for the TOPMed Phase 3 and CCDG projects. This included software version HCSHD 3.4.0.38. Clusters per flow cell were targeted at 1500K/mm2 and150 bp paired-end read lengths were generated. Requested coverage per sample was 30X Haploid Coverage (~103 Gb). Library pools were lightly sequenced to precisely predict the performance of each individual library within the pool. Libraries within the pool were rebalanced to produce a second rebalanced library pool that was fine-tuned to produce optimal coverage. Rebalancing was achieved by examining read counts from the initial data, with subsequent pools adjusted based on anticipated coverage levels. This method ensured that <99% of samples met or exceeded the quality control metrics. Each flow cell was evaluated by reviewing %pass filter clusters (>55%), %^{3}Q30 for Read 1 and Read 2 (avg. >75%) and PhiX error rates for Read 1 and Read 2 (avg. <2%).

Read Processing

Data from each run was demultiplexed using bcl2fastq2 (V2.16.0.10) and aligned to build 38 using BWA-MEM 0.7.15. Post processing included the following steps: Samblaster 0.1.2.4 - add mate tags, Samtools merge 1.3.1 - align and tag files, Sambamba 0.6.4 - name sort, and Picard 2.4.1 - mark duplicates. Base recalibration was achieved using GATK 3.6 and files were converted to the CRAM format using Samtools 1.3.1. Variants were called using GATK Haplotypecaller 3.5. Genotype analysis was accomplished using a custom MGI program version 1.0.

Sequence Data QC

Before transfer of the data to the IRC, Picard 2.4.1 was used to generate statistics to check the quality of each sample. The flagstat output file was generated using Samtools 1.3.1. Contamination and genotype identity were determined using VerifybamID 2 and VerifyBamID 1.1.3. The total number of aligned, non-duplicate bases with BaseQ20 score was generated using bamUtil 1.0.13.

The data generated for each sample met the following TOPMed and CCDG metrics: haploid coverage (mean mappable coverage) must be above 30x, FREEMIX Alpha Score (sequence-only estimate of contamination) must be below 0.01, genotyping CHIPMIX score must be below 0.01, 90% of genome covered at 20x, 95% at 10x, minimum of 86 x 10^9 mapped, aligned bases with Q20 or higher, with minimum coverage calculated after duplicate removal.

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.

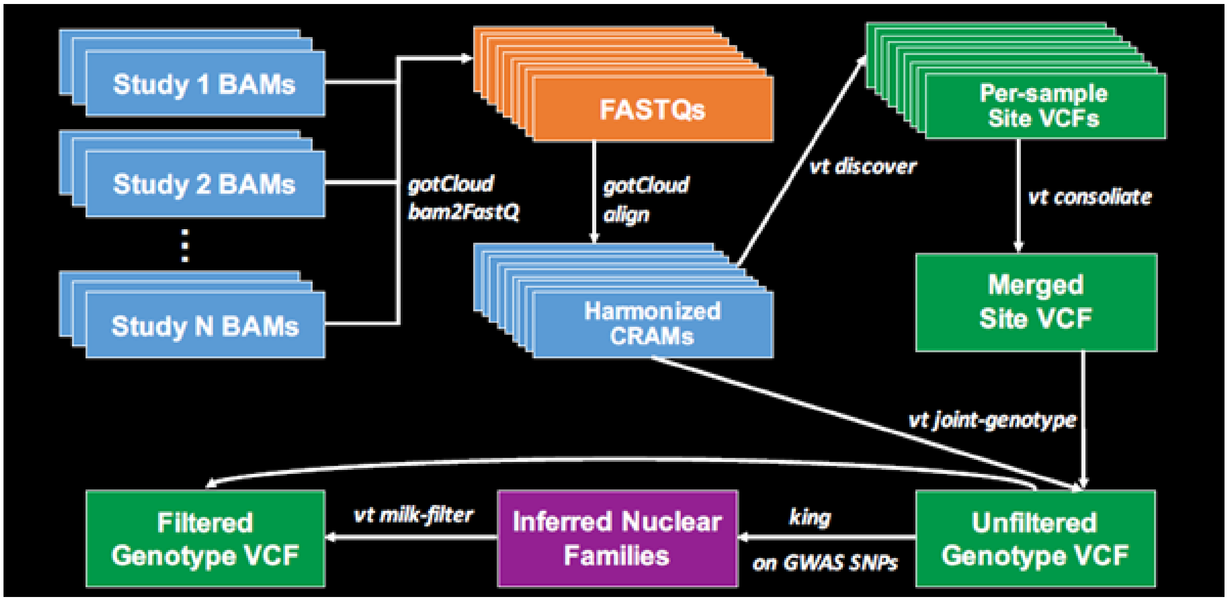

Figure 1 : Schematic view of IRC alignment and variant calling pipeline

Harmonization of Read Alignments

Starting with data freeze 5, all TOPMed sequence data are mapped to the GRCh38 human genome reference sequence in a manner consistent with the joint CCDG / TOPMed functionally equivalent read mapping pipeline described in (Regier, A. et al. 2018. Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. Nature Comm, v.9, n.1, art.4038, DOI: 10.1038/s41467-018-06159-4, PMID: 30279509). TOPMed samples sequenced in phases 1, 2 and 3 were remapped using the Michigan implementation of the mapping pipeline. TOPMed samples sequenced in phases 4 and 5, and those sequenced in the NHGRI CCDG program, were mapped using each sequencing center's implementation of the functionally equivalent mapping pipeline.

Sequence data are received from each sequencing center in the form of .bam or .cram files mapped to the 1000 Genomes hs37d5 build 37 or GRCh38DH build 38 human genome reference sequences. File transfer is 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 validates the md5 checksum, indexes each .bam file using 'samtools index' and uses 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 add ''NWD'' DNA sample identifiers to the read group header lines (Illumina) and convert from UCSC to Ensembl chromosome names (Illumina and Macrogen) using 'samtools reheader'. In-house scripts are used to add read group tags as needed to legacy Illumina sequencing data from 2012-2013.

The two sequence quality criteria used in freeze 8 in order to pass sequence data on for joint variant discovery and genotyping are: estimated DNA sample contamination below 10%, and 95% or more of the genome covered to 10x or greater. DNA sample contamination is estimated from the sequencing center read mapping using an updated version of the verifyBamId software (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).

Descriptions of the IRC's local and standard software tools are available from:

http://genome.sph.umich.edu/wiki/BamUtils
http://genome.sph.umich.edu/wiki/GotCloud
http://www.htslib.org (samtools)
https://github.com/lh3/bwa (bwa, current)
https://github.com/CCDG/Pipeline-Standardization/blob/master/PipelineStandard.md

Software sources:

https://github.com/statgen/bamUtil/releases/tags/v1.0.14
https://github.com/statgen/qplot
https://github.com/Griffan/verifyBamID/releases/tags/1.0.1
https://github.com/samtools/samtools/archive/1.3.1.zip
https://github.com/lh3/bwa (source code)
https://github.com/lh3/bwa/tree/master/bwakit

GRCh38 human genome reference source:

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

Access to sequence data

Copies of individual level sequence data for each study participant are stored on both Google and Amazon clouds. Access involves an approved dbGaP data access request (DAR) and is mediated by the NCBI Sequence Data Delivery Pilot ("SDDP") mechanism. This uses 'fusera' software running on the user's cloud instance to handle authentication and authorization with dbGaP. It provides read access to sequence data for one or more TOPMed (or other) samples as .cram files (and associated .crai index files) within a fuse virtual file system mounted on the cloud computing instance. Samples are identified by "SRR" run accession numbers assigned in the NCBI Sequence Read Archive database and shown under each study's phs number in the SRA Run Selector (https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi). The fusera software is limited to running on Google or Amazon cloud instances to avoid incurring data egress charges.

Fusera: https://github.com/mitre/fusera
Docker: https://hub.docker.com/r/statgen/statgen-tools

Variant Discovery and Genotype Calling

Overview

The freeze 8 genotype call set is 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 and Gonçalo Abecasis). The software tools in this version of the pipeline are available on github at https://github.com/statgen/topmed_variant_calling. The following description refers to specific components of the pipeline. These variant calling software tools are under continuous development; updated versions can be accessed at http://github.com/atks/vt or http://github.com/hyunminkang/apigenome.

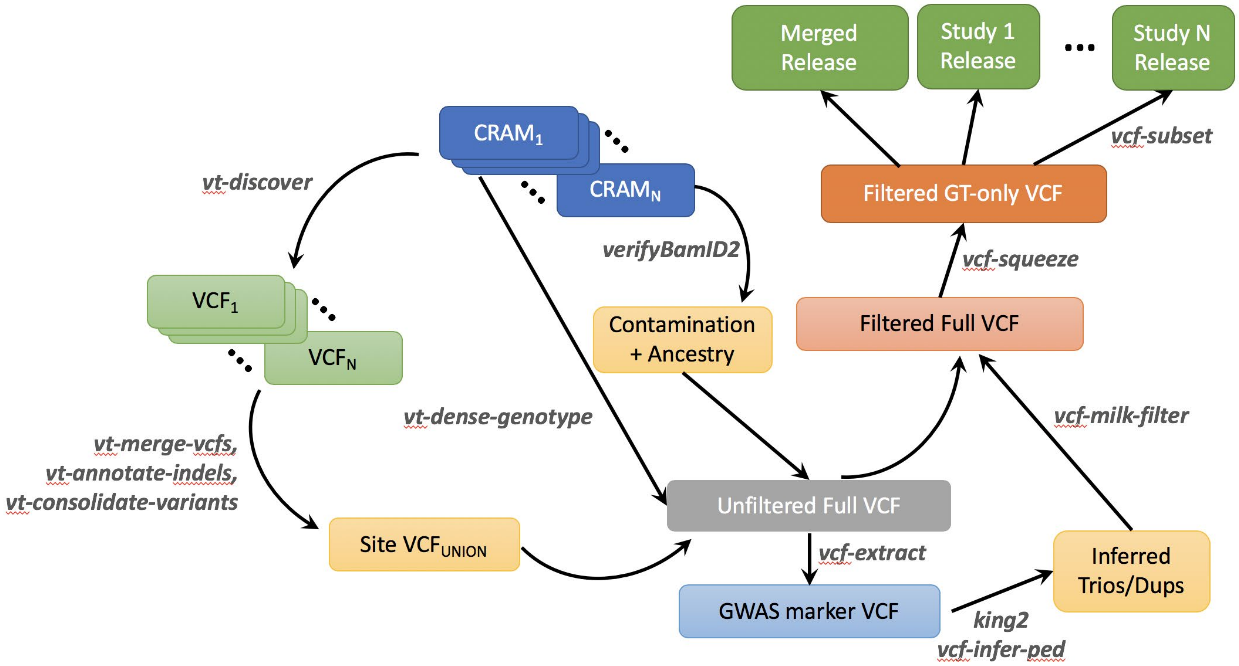

Figure 2 : Outline of TOPMed Freeze 8 Variant Calling Pipeline

Outline of the variant calling procedure

The `GotCloud` pipeline detects variant sites and calls genotypes for a list of samples with aligned sequence reads. Specifically, the pipeline for freeze 8 consists of the following six key steps (see also Figure 2). These procedures have been integrated into the current release of the GotCloud software package at https://github.com/statgen/topmed_variant_calling.

1. **Sample quality control** : For each sequenced genome, genetic ancestry and DNA sequence contamination are estimated by the `cramore cram-verify-bam` software tool. In addition, the biological sex of each sequenced genome is inferred from the relative depth of X and Y chromosomes compared to the autosomal chromosomes, using the software tool `cramore vcf-normalized-depth`.
2. **Variant detection** : For each sequenced genome (in BAM/CRAMs), candidate variants are detected by the `vt discover2` software tool, separately for each chromosome. The representation of indels is normalized by the `vt normalize` algorithm.
3. **Variant consolidation** : For each chromosome, called variant sites are merged across all samples, accounting for overlap of variants between samples, 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, in 10Mb 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 1. These per-batch genotypes are merged across all batches in 100 kb regions using the `cramore paste-vcf-calls` software tool, producing merged and unfiltered genotypes. The

estimated genetic ancestry of each individual is 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 are extracted using `cramore vcf-squeeze` and `cramore vcf-extract` tools. These genotypes and the inferred sex from step 1 are used together to infer a pedigree consisting of duplicate individuals and nuclear families using the `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.

See https://github.com/statgen/topmed_variant_calling/blob/master/README.md for detailed step-by-step instructions to run the variant calling pipeline on an example data set of 1000 Genomes samples. Some details for individual steps are discussed below.

Variant Detection

Variant detection from each sequenced (and aligned) genome is 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 later 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 the `vt normalize` software tool.

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 the `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 http://www.ncbi.nlm.nih.gov/pubmed/25884587.

This 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 are merged across all samples.

1. The non-reference alleles normalized by `vt normalize` algorithm are merged across the samples, and unique alleles are represented as biallelic candidate variants. The algorithm is published at http://www.ncbi.nlm.nih.gov/pubmed/25701572.
2. For alleles which overlap with other SNPs or indels, `overlap_snp` and `overlap_indel` tags are added in the FILTER column of the corresponding variant.
3. If there are tandem repeats with 2 or more repeats with total repeat length of 6bp or longer, the variant is annotated as a potential VNTR (Variable Number of Tandem Repeats), and `overlap_vntr` tags are added to any variant overlapping with the repeat tract of the putative VNTR.

Variant Genotyping and Feature Collection

The genotyping step iterates over all of the merged variant sites and over all 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 genotype likelihoods 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(FIBC_I)
- 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 includes adjustment for potential contamination. It uses an adjusted genotype likelihood similar to the published method https://github.com/hyunminkang/cleancall, but does not use estimated population allele frequency for the sake of computational efficiency.

It conservatively models the probability of observing a non-reference read given a homozygous reference genotype as half of the estimated contamination level (or 1%, whichever is greater). The probability of observing a reference read given a homozygous non-reference genotype is calculated in a similar way. This adjustment calls a heterozygous genotype more conservatively when the numbers of 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 a homozygous reference genotype while the original method, ignoring potential contamination, would call a heterozygous genotype. This adjustment improves the genotype quality for contaminated samples and reduces genotype errors by several fold.

Variant Filtering

The variant filtering in TOPMed Freeze 8 is performed by (1) first calculating Mendelian consistency scores using known familial relatedness and duplicates, and (2) training a Support Vector Machine (SVM) classifier between known variant sites (positive labels) and Mendelian inconsistent variants (negative labels).

Known variant sites are SNPs found to be polymorphic either in the 1000 Genomes Omni2.5 array or in HapMap 3.3, with additional evidence of being polymorphic in the sequenced samples. Negative labels are defined when the Bayes Factor for Mendelian consistency, quantified as `Pr(Reads | HWE, Pedigree) / Pr(Reads | HWD, no Pedigree)`, is less than 0.001. In addition, a variant is marked with a negative label if 10% or more of families or pairs of duplicate samples (and more than 3 trios or duplicate pairs) show Mendelian inconsistency within families or genotype discordance between duplicate samples. Variants eligible to be marked with both positive and negative labels are discarded from the labels. The SVM scores trained and predicted by the libSVM software tool are annotated in the VCF file.

Two additional hard filters are applied. (1) Excess heterozygosity filter (`EXHET`), when the Hardy-Weinberg disequilibrium p-value is less than 1e-6 in the direction of excess heterozygosity after accounting for population structure. An additional ~3,900 variants are filtered out by this filter. (2) Mendelian discordance filter (DISC), when 5% or more of families (and more than 2 trios or duplicate pairs) show Mendelian inconsistency or genotype discordance. An additional ~370,000 variants are 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. An additional level of per-genotype QC is available in "minDP10" genotype files, since these set to missing any individual genotype based on fewer than 10 covering sequence reads. Phased haplotypes are produced by statistical phasing with Eagle 2.4 (Dec 13, 2017). Phasing is done in 1 Mb chunks (with 100 Kb overlap between chunks) for variants which pass all filters, starting with minDP10 genotypes to restrict to high quality genotypes. Phasing re-imputes any missing genotypes.

Eagle: https://github.com/poruloh/Eagle

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). 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 and/or pedigree checks. We also identified a small number of apparent sex chromosome aneuploidies (e.g., XXY, XXX, XYY). If applicable to a study, these are annotated in a file accompanying the genotypes, with flags for “Xchr.anom” and “Ychr.anom”.

Concordance between prior SNP array genotypes and WGS-derived genotypes

Prior genome-wide SNP array data are available for 42 of the 72 accessions to be released starting in 2019.

For some 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 other accessions, a subset of autosomal SNPs with MAF > 5% on a genome-wide array were used (with variants selected to overlap fingerprints if possible). Some accessions had a combination of fingerprint and other array data. For either fingerprint and/or other 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 release or resolved as sample switches using pedigree and/or sex-mismatch results.

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

Comparisons of observed and expected relatedness from pedigrees

Kinship coefficients (KCs) were estimated for pairs of individuals using ~638k single nucleotide variants that are autosomal, MAF >1%, missing call rate <1%, and pruned to have low linkage disequilibrium ($r^2<0.1$) with one another. The estimation procedure used KING IBD segment inference (http://people.virginia.edu/~wc9c/KING/manual.html#IBDSEG). 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, phs001416, phs001466, phs001515, phs001607, and phs001726). 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.