All Of Us Research Program

Genomic Research Data Quality Report

Overview

This document details the *All of Us* Genome Centers (GC) and Data and Research Center (DRC) quality control (QC) steps for genomic data in the research pipeline. This pipeline removes or flags samples and variants in the genomic data that fail quality thresholds. We apply these QC steps in the research pipeline before we release the genomic data for research use. We, the *All Of Us* Data and Research Center (DRC), only describe QC processes that are performed analytically (i.e., after the sample has been genotyped and sequenced). All descriptions and results are limited to the Q2 2022 release made available in the Researcher Workbench June 22, 2022, which contains 165,127 array samples and 98,590 whole genome sequencing (WGS) samples. The samples in the genomic data correspond to the *All of Us* Curated Data Repository (CDR) release C2022Q2R2. These pipelines are automated unless otherwise noted. This document covers all genomic data types made available to researchers at this time including small variants (SNPs and Indels) for arrays and short-read whole genome sequencing (WGS).

Audience: This document is intended for researchers using, or considering the use of, the genomic data in the Researcher Workbench (RWB). This document assumes knowledge of sequencing, genotype arrays, common genomic data QC approaches, and the variant file formats released in *All of Us*. We recommend that at a minimum researchers read the **Known Issues** section below, even if they are not as concerned with the QC process.

Notes:

- Details of the processing (e.g., algorithms) are out of scope for this document.
- The QC process for extracting and cataloging DNA samples is out of scope for this document, since this process happens before genotyping and sequencing.
- Failed samples are not reported here unless otherwise noted.
- Raw data and sample lists will be published to the User Support Hub [1] for researchers. This document does not contain locations of the data.
- The genomic data mentioned in this document requires Controlled Tier access to view. To register for access, please go to https://www.researchallofus.org/register/

Executive Summary

On June 22, 2022, the *All of Us* Research Program released the genomic data of 98,590 WGS and 165,127 array samples in the Researcher Workbench (RWB) for use by users registered for Controlled Tier access. Variant calls from both WGS and arrays (over 702M WGS SNP and indel sites; over 1.8M array SNP and Indel sites), raw data (IDAT files for array data and CRAM files for WGS data), and auxiliary files (predicted ancestry, relatedness/kinship scores, functional annotation, and flagged samples) are available in the RWB (access required). Quality control processes, performed both independently and across samples, indicate that these data are ready for general analysis. We suggest researchers, at a minimum, read the Known Issues section below before using the data.

Introduction

All of Us (AoU) is collecting biospecimens and generating genomic data for all participants who have consented among its target of 1,000,000 participants. As the program continues, the DRC will periodically release genomic data - in sync with planned CDR release timelines. This document describes the second release of genomic data to *All of Us* researchers ("Q2 2022 release") made available in the RWB on June 22, 2022, which contains 165,127 array samples and 98,590 WGS samples, from a diverse set of participants (see Appendix A and Appendix K). All of the released samples with genomic data have at least one other data type (e.g., survey data) that can be joined for analysis (see Appendix L). In this document, we describe the QC processes applied to both the genotyping array ("array") and whole genome sequencing data (WGS). We describe which processes were performed at the GCs and which were performed at the DRC (see Appendix M), but for most researchers this demarcation has no practical significance.

We have split the QC into three conceptual areas:

1. Consistency -- The uniformity of protocols at each GC that reduce the probability of batch effects and that normalize the data across GCs.
2. Single Sample QC -- QC processes run for each sample independently. These catch major errors, such as sample swaps or sample contamination.
3. Joint Callset QC (WGS only) -- QC processes executed on the joint callset, which uses information across samples to flag samples and filter variants.

We have also performed data validation experiments, such as replicating GWAS results, but the results are shown in other, upcoming documentation (see User Support Hub [1] and Tutorial Workspaces in the RWB, both require access).

Consistency across Genome Centers

The genome centers (GCs) established a consistent sample and data processing protocol for array and WGS data generation to attenuate the likelihood of batch effects across GCs. Descriptions in this document, for both QC and sample processing, apply to all GCs unless otherwise noted.

Arrays

The GCs generate variant calls (VCFs) that are submitted to the DRC. The GCs use the same lab protocols, scanners, software, and input files:

- GCs generate raw intensity data (.idat) using the same hardware (iSCAN scanners from Illumina) -- These files will still contain biases across GCs.

- GCs normalize the raw intensity data onto the same scale. This process yields a normalization transform for probe intensities, which are one of the inputs for variant calls. This transform takes into account variation across GCs. Each GC will use the derived clusters to normalize their IDAT files and generate variant calls.
- GCs use identical pipelines to generate VCFs -- This includes both identical pipeline versions and input parameters, where applicable. As a result, the VCFs contain the same information, regardless of GC, including metadata about inputs.

See Appendix J for details on the processing of arrays. Raw array data for each sample in IDAT format are available in the Q2 2022 release for analysis in addition to the variant calls.

WGS

The GCs use the same protocol for library construction (PCR Free Kapa HyperPrep), sequencer (NovaSeq 6000), software (DRAGEN v3.4.12), and software configuration. The software produces the metrics that are consumed by the sample QC processes. For more information about the sequencing processes used by the GCs, see previous work [2] and the NIH *All of Us* Research Program's Return of Genetic Results FDA IDE (G200165). The raw WGS reads for analysis are available in the Q2 2022 release in CRAM data format in addition to the variant calls.

Single Sample QC

The processes documented in this section test each sample, independently. If a sample fails this test, then it is excluded from the release and is not reported in this document. These tests detect sample swaps, cross-individual contamination, and sample preparation errors. In some cases, we perform these tests twice for two reasons: 1) to confirm internal consistency between the GCs and the DRC and 2) to mark samples as passing (or failing) QC based on the research pipeline criteria. The single sample QC process accepts a higher contamination rate than the clinical pipeline (0.03 for the research pipeline versus 0.01 for the clinical pipeline), but otherwise uses identical thresholds. The list of specific QC processes and an overview of the results can be found in Table 1.

Our WGS single sample QC uses the same sequencing process described previously [2] and in the NIH *All of Us* Research Program's Return of Genetic Results FDA IDE (G200165). The processes described previously include single sample QC processes that are not described here. The processes in this document focus on downstream analytical QC processes after a sample has been sequenced or genotyped.

For more details about the array single sample QC process, including preparation, see Appendix J.

Table 1 -- Single Sample QC processes

QC process	Data types	Passing criteria	Error modes addressed	Q2 2022 release results

Fingerprint concordance	WGS (uses Arrays)	log-likelihood ratio > -3	-Sample swaps -Large amount of sample contamination	All array and WGS sample pairs are concordant.
Sex concordance	WGS and Arrays	Sex call is concordant with self-reported sex at birth. OR Self-reported sex at birth reported as “Other” or was not reported	-Sample swaps	All array and WGS samples are concordant.
Call rate	Arrays	> 0.98 (> 98%)	-Sample contamination -Sample preparation error	All array samples meet the threshold. Inconsistency across GCs was discovered. See the Call Rate Section and Known Issues #2 We erroneously failed 2994 array samples, which are not included in this release. However, we have included the corresponding WGS samples. See Known Issues #1.
Cross-individual contamination rate	WGS and Arrays	WGS: < 0.03 (< 3%) Arrays: None (Reported only)	Sample contamination from another individual	All WGS samples meet the threshold. For arrays, we only report the contamination rate, but do not filter array samples, since the call rate is a proxy for high levels of contamination. WGS samples with corresponding arrays that have a contamination rate above 10% were not released.
Coverage	WGS	≥ 30x mean coverage ≥ 90% of bases at 20x coverage ≥8e10 aligned Q30 Bases ≥ 95% at 20x in regions of the 59 AoU Hereditary Disease Risk genes (AoUHDR) See Appendix F for more information	-Sample preparation error -Poor sensitivity and precision of variant calling	All WGS samples meet the thresholds.

Fingerprint Concordance

Method

We filter variant calls to 114 sites ("fingerprint") for both the array and WGS variants. We measure the concordance between the array and WGS data, using a log-likelihood ratio (fingerprint LOD) based on reads. We chose the threshold value, -3.0, to split a bimodal distribution (not shown). If the calls are not concordant (i.e., the fingerprint LOD does not meet the threshold), then there has likely been a sample processing error. A detailed description of fingerprint concordance is described in the Genome Analysis Toolkit documentation. [3]

We call the fingerprint concordance using Picard (version 2.23.9) with the following parameters:

Parameter	Value
program name	"CheckFingerprint"
INPUT	The WGS cram to check concordance
REFERENCE_SEQUENCE	"gs://gcp-public-data--broad-references/hg38/v0/Homo_sapiens_assembly38.fasta"
GENOTYPES	VCF from corresponding array file
HAPLOTYPE_MAP	"gs://gcp-public-data--broad-references/hg38/v0/Homo_sapiens_assembly38.haplotype_database.txt"
IGNORE_READ_GROUPS	"true"
SAMPLE_ALIAS	Chipwell barcode from the header of the array file (array file passed in the GENOTYPES parameter)

Note: Quoted parameters are exact values, but quotes were not included in the actual call to the tool.

Results

All samples in the Q2 2022 release passed the fingerprint concordance check. We were able to run fingerprint checks on WGS samples using the arrays, but 2,994 of the corresponding array files were not included in this release (see Known Issue #1)
As seen in Figure 1, the passing samples exceeded the threshold. Fourteen samples had a fingerprint LOD [3] less than 45 and the minimum fingerprint LOD was 13.

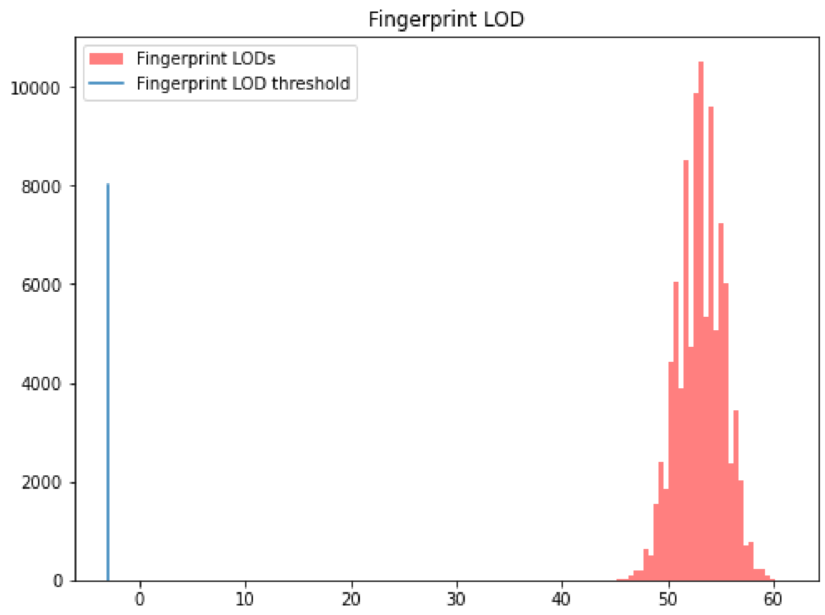


Figure 1 -- Distribution of the Fingerprint LODs for WGS Q2 2022 samples

Sex Concordance

We checked the computed sex against the self-reported sex at birth for concordance (see Appendix H). If the two sources were not concordant, we assumed a potential sample swap, removed the sample, and investigated the source of the swap. If we do not have a “male” or “female” for the sex assigned at birth, because the participant reported it as “Intersex”, “I prefer not to answer”, “none of these fully describe me”, or skipped the question, we passed the sex concordance check for that sample.

WGS

Method

We compare variant and ploidy calls for chromosome X and Y against the self-reported sex assigned at birth for the sample. We check the sex ploidy call (e.g., XY or XX) from the DRAGEN pipeline (v 3.4.12) and use heterozygous chrX variant calls from peddy [4]. If the concordance test fails against either of these calls, the sample fails QC and is not included in the release. If we do not have a “male” or “female” for the sex assigned at birth, because the particpant reported it as “Other” or skipped the question, we will pass sex concordance regardless of the information from peddy and DRAGEN.

DRAGEN invocations include a wide breadth of functionality, including ploidy calls (see Appendix G for the parameters).

The DRAGEN pipeline outputs a single sample VCF, which is primarily used in the clinical pipeline (for individual samples), but we use it for our call to peddy. We call peddy with the following parameters:

Parameter	Value
vcf	Single sample VCF from DRAGEN (hard-filtered)
Pedigree file	We create this file dynamically based on the single sample and its sex call. Please note: This implies that we do not use pedigree information in our peddy call.

Results

Since we catch sex concordance failures before including a sample in the release, all WGS samples in the Q2 2022 release passed a sex concordance check. Note that 1.5% of samples passed the sex concordance check due solely to their answer on the self-reported sex assigned at birth ("I prefer not to answer", "none of these fully describe me", "Intersex", or skipped the question). See Appendix H for a full breakdown of self-reported sex assigned at birth.

Array

Method

We call the gencall tool [5] v3.0.0 to make a call on the sex of the sample. We use the Picard 2.26.0 tool, CollectArraysVariantCallingMetrics [6], to perform the actual concordance check against the self-reported sex assigned at birth. If we do not have a "male" or "female" for the sex assigned at birth because the participant reported it as "Other", "Intersex", or skipped the question, we will pass sex concordance regardless of the sex call from the array.

To generate sex calls from the array, we call gencall from the Illumina Array Analysis Platform Genotyping Command Line Interface (iaap-cli):

Parameter	Value	Notes
Tool name	"gencall"	
Manifest file	Bead pool manifest (BPM)	Illumina-supplied file that contains metadata (alleles, mapping information, source, etc.) for all of the probes on the genotyping array.
Cluster file	Cluster file (EGT)	Used for normalization of intensities across GCs
-f	Location of the IDAT (.idat) files	
-i	"1"	Algorithm version

--gender-estimate-call-rate-threshold	-0.1	This effectively disables the sex estimation.

To ensure concordance with the self-reported sex assigned at birth, we call CollectArraysVariantCallingMetrics with the following parameters from the Picard toolkit:

Parameter	Value
Tool name	“CollectArraysVariantCallingMetrics”
INPUT	Array single sample VCF
DBSNP	"gs://gcp-public-data--broad-references/hg38/v0/Homo_sapiens_assembly38.dbsnp138.vcf"

Results

Since we catch sex concordance failures before including a sample in the release, all array samples in the Q2 2022 release passed a sex concordance check. Note that 1.5% of samples passed the sex concordance check due solely to their answer on the self-reported sex assigned at birth (“I prefer not to answer”, “none of these fully describe me”, “Intersex”, or skipped the question). See Appendix H for a full breakdown of self-reported sex assigned at birth.

Call Rate (Array only)

Method

The call rate is the number of successful variant calls divided by the number of probes. We invoke the gencall tool [5] v3.0.0, as described above in *Sex Concordance*, which generates both sex calls and the call rate. We also invoke CollectArraysVariantCallingMetrics with the same parameters to extract the call rate metric from the VCF header.

We applied a threshold of 0.98 to the call rate for inclusion in the Q2 2022 release, but we believe that we were overly-aggressive filtering samples due to an internal inconsistency with call rate methodology applied across GCs (see Known Issue #2).

Results

As seen in Figure 2, we did not include any samples that were below the call rate threshold of 0.98. During the generation of the release, we discovered an inconsistency across GCs in the calculation of call rates. The methodology was updated to make the GCs consistent, but this resulted in two separate call rate populations, as seen in Figure 2. These dual peaks hold for all three GCs, as seen in Figure 3.

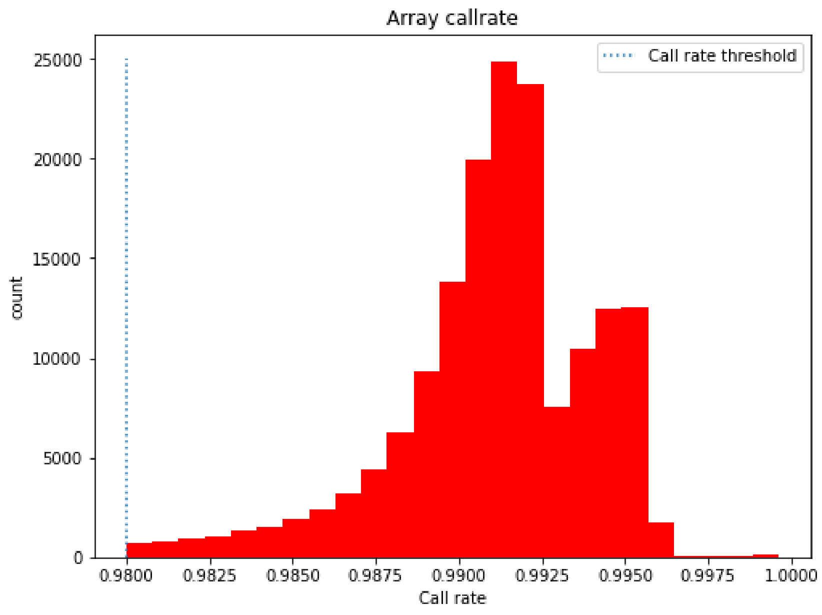


Figure 2 -- Histogram of the array call rate for the Q2 2022 release. Note that a correction in call rate calculation led to two peaks in the histogram.

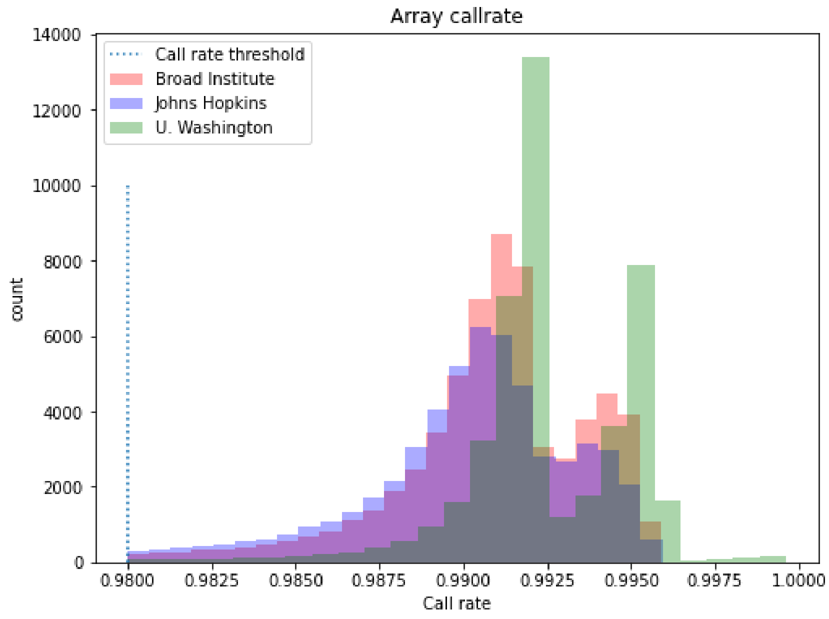


Figure 3 -- Call rate across each GC. Note that the bimodal distribution is seen across centers.

Cross-individual Contamination Rate

For all samples, we estimate the proportion of data coming from an individual other than the one being processed, referred to as the contamination rate. We follow two separate processes for WGS and arrays. Samples can only fail a contamination rate check for WGS. For arrays, as the contamination rate increases, we expect a lower call rate. We fail array samples for a call rate that does not meet the threshold.

WGS

Method

We estimate the percent contamination from another individual by counting the number of reads at common homozygous alternate SNP sites. If there is a small amount of cross-individual contamination, we expect to see small numbers of reads supporting SNPs at these sites. We determine the percentage of the sample that may have come from a different individual using VerifyBamID2 [7], and the DRAGEN 3.4.12 pipeline. Contamination rate is a float value from 0.0 to 1.0, which represents 0 to 100%.

DRAGEN invocations include a wide breadth of functionality, including contamination estimates (see Appendix G for the parameters).

We use the following parameters for VerifyBamID2:

Parameter	Value
NumPC	“4”
BamFile	WGS cram file
Reference	“gs://gcp-public-data--broad-references/hg38/v0/Homo_sapiens_assembly38.fasta”
UDPath	“gs://gcp-public-data--broad-references/hg38/v0/contamination-resources/1000g/1000g.phase3.100k.b38.vcf.gz.dat.UD”
BedPath	“gs://gcp-public-data--broad-references/hg38/v0/contamination-resources/1000g/1000g.phase3.100k.b38.vcf.gz.dat.bed”
MeanPath	“gs://gcp-public-data--broad-references/hg38/v0/contamination-resources/1000g/1000g.phase3.100k.b38.vcf.gz.dat.mu”
Verbose	specified

Results

We did not include any samples with a contamination larger than 0.018 and only three samples greater than 0.015. See Figure 4 for the frequency of contamination estimates for samples in the Q2 2022 release.

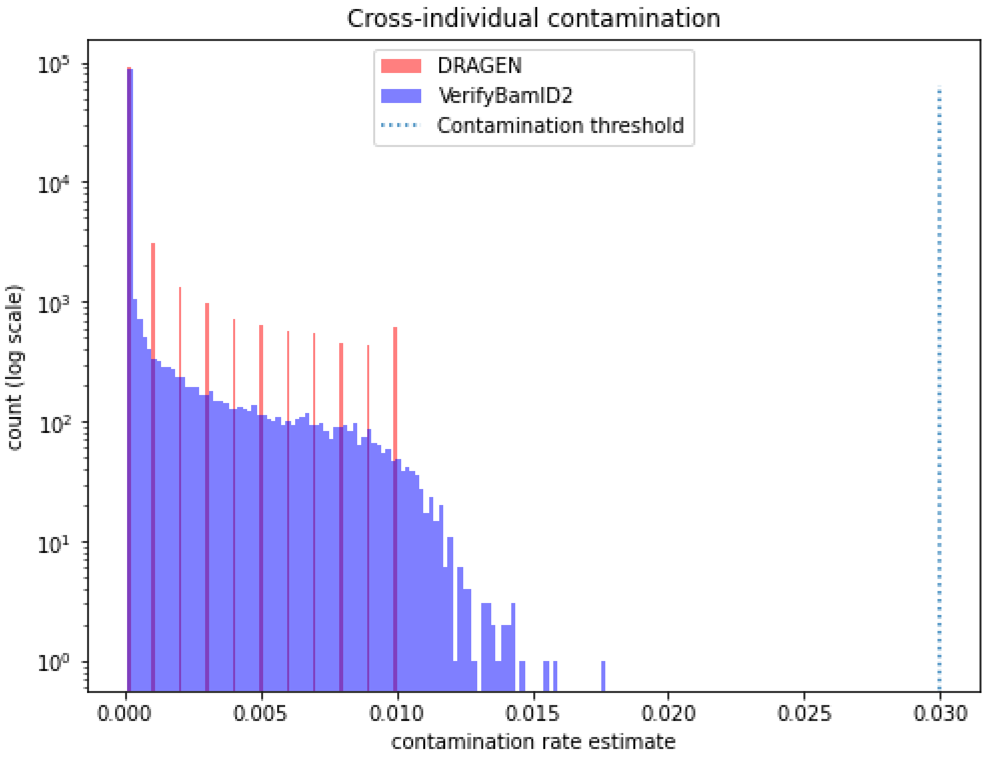


Figure 4 -- WGS contamination estimates from both sources (DRAGEN and VerifyBamID2). DRAGEN rounds the contamination estimate to three decimal places. Note the log scale of the counts (y-axis). Over 88.1% and 90.6% of WGS samples had contamination estimates lower than 1e-4 by VerifyBamID2 and DRAGEN, respectively. Any samples above the contamination threshold are not included in the Q2 2022 release.

Array

Method

We use BAFRegress [8] to estimate the contamination rate in our array data. We do not use the cross-individual contamination rate to filter array samples, and we do not process the corresponding WGS aliquots for any array sample with a contamination greater than 10%. We filter samples based on the call rate, which is a proxy for contamination and other errors, such as sample preparation errors. Note that most samples with a contamination rate greater than 10% will also not meet the call rate threshold.

We extract allele frequency information from the array VCF and convert it into the file format expected by BAFRegress. We then invoke BAFRegress with the following parameters:

Parameter	Value
task	“estimate”
freqfile	Allele frequency information for all sites, which was extracted from the single sample array VCF.

Results

We estimated the contamination rate below 0.11 for all array samples. As the contamination rate increased, we did see a small decrease in the call rate (see Figure 5). Of the 165,127 array samples, 99% had an estimated contamination rate below 3.5%. 159,853 array samples (96.8%) had a contamination rate less than 3% and 154,917 (93.8%) had a contamination rate less than 1%.

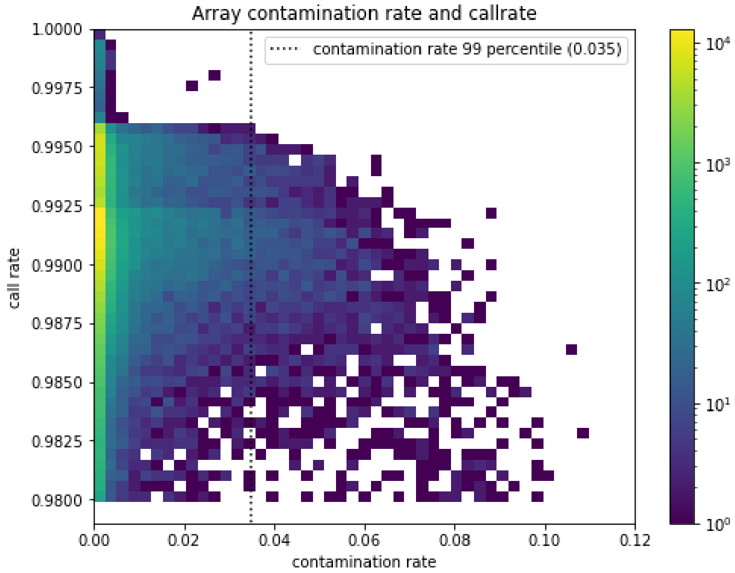


Figure 5 -- Histogram of the array contamination rate estimates vs call rate. As the contamination rate increases, the call rate decreases.

Coverage (WGS only)

Method

Coverage is defined as the number of reads covering the bases of the genome. Maintaining coverage is important for consistent statistical power and accurate variant calling. We apply several thresholds (summarized from the FDA IDE (G200165)):

- Mean coverage (threshold ≥30x) - This is the mean number of overlapping reads at every targeted base of the genome. Accuracy steadily decreases as mean coverage decreases, with a rapid decrease below 20x coverage, supporting a stringent threshold selection of a minimum of 30x.
- Genome coverage (threshold ≥90% at 20x) - Accuracy steadily decreases as the percent of bases with at least 20x coverage drops. Drop-off of performance is initially gradual, supporting a threshold of 90%.
- *All of Us* Hereditary Disease Risk gene (AoUHDR) coverage (threshold ≥95% at 20x) - For clinically relevant areas of the genome, we insist on higher mean coverage to ensure

a higher calling accuracy. As we reduce the coverage in the AoUHDR region, the reduction in performance is slow initially but increases rapidly below 40%, showing that the threshold of 95% is conservative.

- Aligned Q30 bases (threshold ≥8e10) - All bases in the sequencing reads get a quality assignment, which is phred scaled (Q30 → probability of error is 0.001) [9]. As lower base quality counts increase, we see a reduction in accuracy with an inflection point starting around 6e10.

Result

As seen in Figure 6, all WGS samples exceed the thresholds that we set as part of the research pipeline. We had 107 samples with mean coverage greater than 70x. None of these samples were flagged in our joint callset QC.

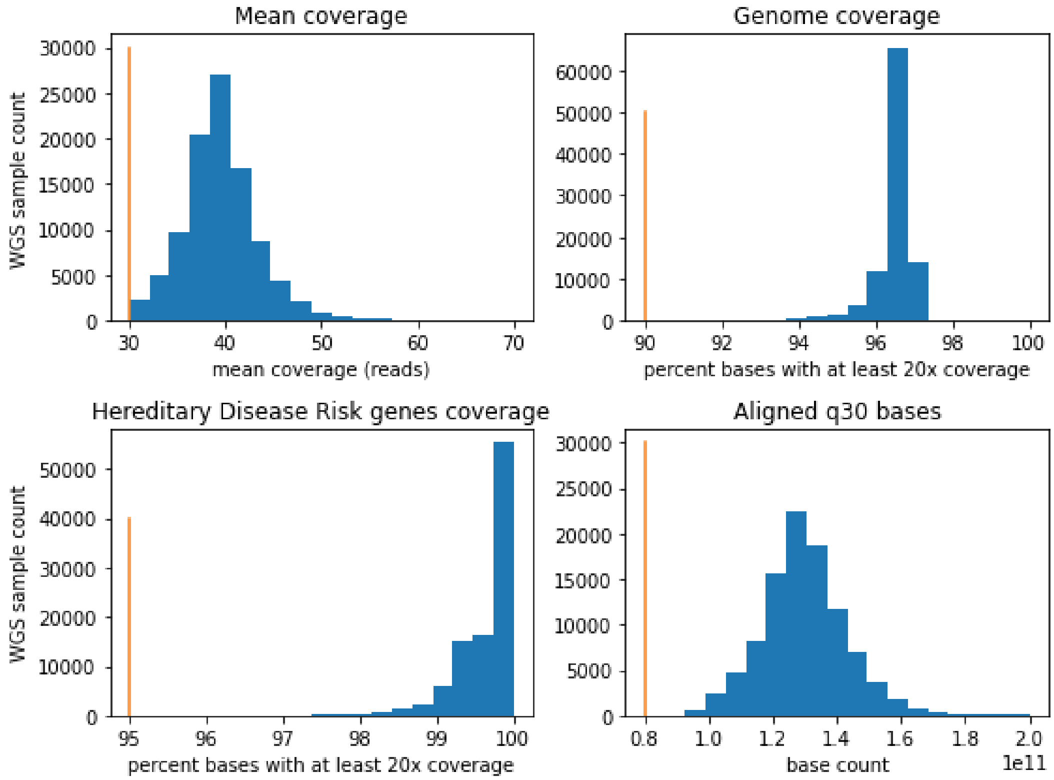


Figure 6 -- Coverage metrics for the Q2 2022 release WGS samples. The orange line is the threshold for each metric. There are 107 samples (0.1%), with mean coverage greater than 70x, that are not included in the mean coverage (upper left) nor aligned q30 bases (lower right) plots. As expected, these samples were outliers in the number of aligned q30 bases (i.e., higher base count than samples with lower mean coverage).

Joint Callset QC (WGS only)

We deliver our WGS variants as a joint callset [10]. We perform QC on joint callsets and make the output accessible to researchers in the RWB. Please note that the QC steps described here apply during creation of the WGS joint callset. These QC steps are not run on individual samples (e.g., GVCFs), though we flag individual samples based on these QC metrics. The list of flagged samples and other auxiliary information, such as ancestry predictions, is available through the User Support Hub [1]. The joint callset QC process is similar to that of gnomAD 3.1 [11], though not exactly the same. We have described our process here and it is summarized in Table 2.

Table 2 -- Joint callset QC summary

QC process	Variant/ sample?	Error modes addressed	Notes
Hard Thresholds	sample	Extremely noisy samples	No samples flagged.
Population Outlier	sample	Noisy samples	156 samples flagged (0.09%). Based on regressing out the PCAs from callset metrics, such as snp_count.
Hard Threshold Filters	variant	Artifacts that cannot be detected in a single sample	This has a simple implementation with high precision, which saves compute for downstream variant filtering. 45,423,717 were filtered 657,151,220 were not filtered
Allele-Specific VariantQualityScore Recalibration (AS-VQSR)	variant	Artifacts that cannot be detected in a single sample	See [12].
Sensitivity and Precision Evaluation	both	Poor variant detection	See Appendix D for a list of samples.
Auxiliary processes			
Ancestry	sample	Flagging sample outliers and allows calculation of population level metrics, such as allele frequency (AF).	Error rate from holdout set (incl. Other): 0.046 Error rate from holdout set (not incl. Other): 0.009 Concordance vs self-reported: 0.877 See Appendix A. Number of independent, bi-allelic sites ("high-quality sites") used: 56695 See Appendix B.
Relatedness and maximal independent set of samples	sample	Related samples, which confound analyses	4846 related pairs and 4069 samples in the maximal independent set. See Appendix C. This process produces a list of the sample pairs with kinship score, calculated by Hail [13]. No samples are removed from the callset, but this allows researchers to easily remove a minimal set of samples to eliminate related samples in the callset.

Method

Below is the list, in order, of the steps to perform the joint callset QC in the Q2 2022 release:

1. Sample hard threshold
2. Sample population outlier
3. Variant hard threshold
4. Allele-Specific Variant Quality Score Recalibration (AS-VQSR) [Filtering]
5. Sensitivity and precision evaluation

The first two steps flag samples ("Sample QC"). The filtering steps (Variant Hard Filtering and AS-VQSR) apply to variants in the joint callset ("Variant QC"). We then measure the sensitivity and precision of the joint callset.

Sample QC

We flagged samples as failing QC, rather than removing them from the callset, since we could not validate whether samples (especially population outliers) were problematic or were just a part of a poorly-sampled ancestry. Flagged samples will be published in a list to researchers through the User Support Hub [1]. These pipelines will flag samples based on the data from the entire joint callset. Therefore, sample-level QC (e.g., contamination) is handled upstream from the process described here. Sample QC is performed before Variant QC (e.g., Sample QC happens before AS-VQSR)

Hard Threshold Flagging

We believe that some samples will have strong erroneous signals. We flag these from the joint callset as an initial step. The criteria for being eliminated as "obviously erroneous" will be:

- number of SNPs: < 2.4M and > 5.0M
- number of variants not present in gnomAD 3.1: > 100k
- heterozygous to homozygous ratio (SNPs and Indel separately): > 3.3

We calculated all metrics using autosomal territory only.

We did not flag any samples for failing hard thresholds.

Population Outlier Flagging

We regressed out sixteen principal component features computed as part of ancestry prediction (see Appendix A) and used the residuals to determine the outliers. We define outlier samples as being eight median absolute deviations (MADs) away from the median residual in any of the following metrics:

i. number of deletions
ii. number of insertions
iii. number of SNPs
iv. number of variants not present in gnomAD 3.1

v. insertion : deletion ratio
vi. transition : transversion (TiTv) ratio
vii. heterozygous to homozygous ratio (SNPs and Indel separately)

We flagged 156 (0.09%) samples as outliers based on at least one of the above criteria (See Table 3 for details). Plots of the first principal components against these eight metrics can be found in Appendix I.

Table 3 -- Population outlier sample counts

Metric(s) considered	Flagged sample count
Indel heterozygous to homozygous ratio	63
Variants not present in gnomAD 3.1 count	43
Indel heterozygous to homozygous ratio + SNP count	12
Deletion count + Indel heterozygous to homozygous ratio + Insertion count + SNP count	10
Indel heterozygous to homozygous ratio + SNP heterozygous to homozygous ratio	8
Deletion count + Indel heterozygous to homozygous ratio + Insertion count + SNP count + SNP heterozygous to homozygous ratio	4
Ti/Tv ratio + Variants not present in gnomAD 3.1 count	3
Deletion count + SNP count	3
SNP heterozygous to homozygous ratio	3
Deletion count + Insertion count + SNP count	2
Deletion count + Indel heterozygous to homozygous ratio + SNP count	2
Indel heterozygous to homozygous ratio + SNP count + SNP heterozygous to homozygous ratio	2
SNP count	1
Total	156

Variant QC

These processes will flag specific variants from a callset. Filtered variants will be included in cohorts, both the entire callset and cohorts generated using the Cohort Builder. For example, if a cohort was exported to VCF, the variant will appear as filtered in the VCF filter field (“FILTER”).

Hard Threshold Filters

If a variant does not meet the following criteria, it will be filtered (i.e., a value will appear in the FILTER field of VCFs and Hail MatrixTables (MT)):

- No high-quality genotype (GQ>=20, DP>=10, and AB>=0.2 for heterozygotes) called for the variant.
  - Allele Balance (AB) is calculated for each heterozygous variant as the number of bases supporting the least-represented allele over the total number of base observations. In other words, min(AD)/DP for diploid GTs.
  - Filter field value: NO_HQ_GENOTYPES
- ExcessHet < 54.69
  - ExcessHet is a phred-scaled p-value. We cutoff of anything more extreme than a z-score of -4.5 (p-value of 3.4e-06), which phred-scaled is 54.69
  - Filter field value: ExcessHet
- QUAL score is too low (lower than 60 for SNPs;69 for Indels)
  - QUAL tells you how confident we are that there is some kind of variation at a given site. The variation may be present in one or more samples.
  - Filter field value: LowQual

Unfiltered variants will have "." or "PASS" in the FILTER field in the WGS joint callset VCFs and Hail MT. We recommend that researchers do not include sites that were filtered in their analyses.

The variant counts can be found in Table 4.

Table 4 -- Hard threshold filter variant counts

Filters	Numbers
None	657151220
'NO_HQ_GENOTYPES'	23282504
'NO_HQ_GENOTYPES', 'LowQual'	18835182
'LowQual'	2731900
'ExcessHet'	572725
'NO_HQ_GENOTYPES', 'ExcessHet'	1406

Allele-Specific VariantQualityScoreRecalibration

As part of the joint calling, we will filter variants with Allele-Specific Variant Quality Score Recalibration (AS-VQSR or VQSR) [12]. This filtering technique uses machine learning to identify variants across samples that are likely artifacts. We used the following annotations as features for training:

- Variant Confidence/Quality by Depth (AS_QD)
- Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities (AS_MQRankSum)
- Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias (AS_ReadPosRankSum)
- Phred-scaled p-value using Fisher's exact test to detect strand bias (AS_FS)
- RMS Mapping Quality of reference vs alt reads (AS_MQ) [SNPs only]
- Symmetric Odds Ratio of 2x2 contingency table to detect strand bias (AS_SOR)

We used the default training sets as described in the GATK documentation [14], except that we use one additional source of training data (Axiom) for indels. Each training set is assigned a flag whether it is representative of true sites or whether we use the sites for training and also assigned an initial prior likelihood score. Details of these parameters can be found in the GATK documentation [12,14], and the sites can be found as public resource downloads for the GATK [15]. We have reprinted the training resource list below for clarity, including the documentation from the GATK at the time of this writing:

- SNP training sites:
    - Omni -- This resource is a set of polymorphic SNP sites produced by the Omni genotyping array [16]. VQSR will consider that the variants in this resource are representative of true sites (truth=true), and will use them to train the recalibration model (training=true). The prior likelihood we assign to these variants is Q12 (93.69%).
    - HapMap [17] -- This resource is a SNP callset that has been validated to a very high degree of confidence. VQSR will consider that the variants in this resource are representative of true sites (truth=true) and will use them to train the recalibration model (training=true). We will also use these sites later on to choose a threshold for filtering variants based on sensitivity to truth sites. The prior likelihood we assign to these variants is Q15 (96.84%).
    - *1000G* [18] -- This resource is a set of high-confidence SNP sites produced by the 1000 Genomes Project. VQSR will consider that the variants in this resource may contain true variants as well as false positives (truth=false) and will use them to train the recalibration model (training=true). The prior likelihood we assign to these variants is Q10 (90%).
- Indels:
    - Mills [19] -- This resource is an Indel callset that has been validated to a high degree of confidence. VQSR will consider that the variants in this resource are representative of true sites (truth=true) and will use them to train the recalibration model (training=true). The prior likelihood we assign to these variants is Q12 (93.69%).

- Axiom (1000G) -- This resource is an Indel callset based on the Affymetrix Axiom array on 1000 Genomes Project samples [18]. VQSR will consider that the variants in this resource may contain true variants as well as false positives (truth=false) and will use them to train the recalibration model (training=true) The prior likelihood we assign to these variants is Q10 (90%).

Sensitivity and Precision Evaluation

In the callset, we included four well-characterized control samples (four Genomes-in-a-Bottle samples (GiaB) [20] from HapMap [17] and Personal Genome Project; see Appendix D), which we can use to determine sensitivity and precision. The samples were sequenced with the same protocol as AoU. These samples do not appear in any user data (e.g., cohorts built using the RWB).

We use the high confidence calling region, defined by GiaB v4.2.1, as the source of ground truth. In order to be called a true positive, a variant must match the chromosome, position, reference allele, and alternate allele. In cases of sites with multiple alternate alleles, each alternate allele is considered separately. Sensitivity and precision results can be seen in Table 5.

Table 5 -- Sensitivity and precision measurements for control samples using the AoU sequencing protocol

Variant type	Sample	Sensitivity	Precision
SNV	HG-001	0.994	>0.999
	HG-003	0.987	>0.999
	HG-004	0.987	>0.999
	HG-005	0.987	>0.999
Indel	HG-001	0.971	0.996
	HG-003	0.970	0.997
	HG-004	0.972	0.998
	HG-005	0.980	0.999

Known Issues

The issues below apply to the Q2 2022 release genomic data (arrays, WGS, and auxiliary data). These will be addressed in the next callset release (ETA 2022), unless stated otherwise. We have provided suggested actions for researchers to workaround the issue. Sample lists relevant to these issues can be found in the User Support Hub [1].

For known issues that existed in the previous releases, we have updated the text to represent changes (e.g., different sample counts).

1. WGS samples are not a strict subset of the array samples

- Affects:
  - WGS joint callset VCFs
  - WGS joint callset Hail MatrixTable (MT)
- Suggested action:
  - If your analysis explicitly involves cross-analyzing WGS samples and the corresponding arrays: Remove the 2,994 affected WGS files from your analysis.
  - Otherwise: No action
- Description: As described in Known Issue #2 below, the array data are missing 2,994 participants that are included in the WGS samples. The array samples were removed for having a low call rate (under 0.98), but this was due to an inconsistency between the call rate tools being used by the GCs and the DRC. We were still able to use these arrays in our WGS array fingerprint concordance QC step. Once the inconsistency is corrected, we believe that these samples will be above the call rate threshold. Note that none of the corresponding array samples had a call rate below 0.974, even when using the most pessimistic estimate, and none failed any other QC check for arrays.
  - The sample list (2,994 (3.0%) WGS samples) will be provided through the User Support Hub.
- Remediation: We are addressing this in two ways:
  - We will be reprocessing all array data that are part of the Q2 2022 release. As part of this effort, we will be synchronizing the way call rates are calculated. For all future callsets, this will further reduce the possibility of having internal inconsistencies over which samples should be included.
  - Once we have consistent callset calculation, we will implement an automated process which will disallow a WGS sample to be included in the joint callset without a corresponding array that passes the single sample QC.

2. Extraneous array samples were failed due to inconsistency of call rate calculations and are missing from the array data

- Affects:
  - Array VCFs
  - Array Hail MT
  - Array PLINK bed/bim/fam
- Suggested action: None. We will provide the sample list of WGS samples without corresponding arrays in the RWB.

- Description: We failed 2,994 arrays with corresponding WGS for not meeting the call rate threshold of 0.98 (see Table 1), even though these passed clinical call rate QC at the GCs. These arrays did not fail any other single sample QC check, but are not included in the Q2 2022 release array VCFs, Hail MT, or PLINK files. In the single sample QC, the DRC generated lower call rate values than the GCs (see Table 6), which prevented samples from meeting the call rate threshold. The DRC was using a different call rate metric (gencall) than the GCs had agreed to use (GTC); GTC call rate yields a higher value, on average. The DRC was not able to switch GTC call rate, because one GC (Johns Hopkins (JH)) was using an older version of the picard tool that disallows calculation of the GTC call rate. The average difference between the two call rate metrics was 0.004. We believe that this difference in call rates is not large enough to fail the corresponding WGS files outright. Where we had corresponding WGS and an array that only failed the minimum call rate, we used the array file to perform our fingerprint concordance check. All corresponding WGS passed the fingerprint concordance check of the 2,994 arrays failing only the callset rate with corresponding WGS.
- Remediation: As part of the next release, we will be reprocessing all array data that are part of Q2 2022. As part of this effort, we will be synchronizing the way call rates are calculated. This will remove any possibility of having internal inconsistencies over which array samples we include.

Table 6 -- Summary statistics of the DRC and GC array call rates in the Q2 2022 release

Call Rate	DRC	GC
Mean	0.991	0.995
Standard Deviation	0.003	0.003
Minimum	0.980	0.980
25%	0.990	0.994
50%	0.991	0.996
75%	0.993	0.997
Maximum	>0.999	>0.999

3. Ancestry prediction has higher error rates for Middle Eastern ancestry

- Affects:
  - Ancestry predictions
  - Variant Annotation Table (VAT)
  - Public Data Browser

- Suggested Action: When limiting cohorts to samples with computed ancestry of Middle Eastern (“mid”), use the ancestry predictions that do not include “other”. In other words, use the “ancestry_pred” column, instead of “ancestry_pred_other”.
- Description: A paucity of labeled Middle Eastern samples reduced the performance of the random forest classifier. This caused the confidence to dip when predicting ancestry for Middle Eastern samples, which caused a larger proportion of these samples relative to other computed ancestries, to be classified as Other (“oth”). The VAT uses these computed ancestries to generate *All of Us* population (gvs_mid_* and gvs_oth_*) annotations. The genomic data in the public Data Browser are also dependent on the ancestry predictions for populating population information about variants.

  See Table A.2 for details of the ancestry prediction performance
- Remediation: We will investigate other approaches for classification. We will start the updated approach with the next callset.

4. Allelic Depth (AD) has incorrect VCF header line and unconventional format

- Affects:
  - WGS joint callset VCFs
  - WGS joint callset Hail MT
- Suggested Action:
  - If you do not use the AD field, then no action.
  - Otherwise, use the AD field according to the format detailed here. This will likely require a change in analysis code.
- Description:
  - The conventional header for AD is:

    ```
    ##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic
    depths for the ref and alt alleles in the order listed">
    ```

  The AoU callset includes the conventional header in the VCF. However, this does not represent the data. The header should be:

    ```
    ##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic
    depths for the alleles specified in the genotype field (GT), with
    the reference allele always included.  For diploid samples, this
    will have zero (hom ref), two (het ref), or three (het alt)
    values.">
    ```

  Conventionally, the AD field has a value for the reference and each alternate allele (i.e., in the VCF header: Number=R) in the genotypes. We have a different convention in this release, which is a similar (not the same) encoding to Local Allelic Depths in Hail. Please note that we do not use any local allele fields in this release.

  We only provide allelic depth for variants with a non-ref allele (i.e., not a homozygous reference variant). We include the informative read count for all alleles in the genotype field (GT) plus the reference. Therefore, all AD values have zero (homozygous reference), two (heterozygous reference), or three (non-reference heterozygous) numbers. For example, if the genotype is “0/1” and AD is “[10,20]”, then there were 10 reads supporting the reference (GT 0) and 20 reads supporting the first alternate allele (GT 1). If the genotype is 1/3 and AD is “[0, 5, 10]”, then there were no reads supporting the reference, 5 reads supporting the first alternate allele (GT 1), and 10 reads supporting the third alternate allele (GT 3). Table 7 has formatting of the AD field values given examples of alternate allele counts and a corresponding GT.

Table 7 -- *All Of Us* Q2 2022 release AD field examples

Variant Type	Number of alternate alleles	Genotype call (GT)	Resulting Allelic Depth (AD)
Homozygous Reference	<any>	0/0	. (missing)
Heterozygous Reference	1	0/1	[N_0,N_1]
	7	0/4	[N_0,N_4]
Homozygous Alternate	1	1/1	[N_0*,N_1]
	7	3/3	[N_0*,N_3]
Heterozygous Alternate	2	1/2	[N_0*, N_1, N_2]
	7	2/4	[N_0*, N_2, N_4]
No Call	<any>	./.	. (missing)

* - These will usually be zero, since these are read counts of the reference allele in a non-reference variant.

- Remediation (ETA 2022):
  - We will rename the AD field in the AoU WGS joint callset and include a correct header for the renamed field. This will minimize confusion between AD in AoU genomic data and the conventional definition of AD.

5. Extraneous INFO field (AS_YNG) in the WGS data

- Affects:
  - WGS joint callset VCFs
  - WGS joint callset Hail MT
- Suggested Action: Do not include the AS_YNG field in any analyses.
- Description: The WGS joint callset includes AS_YNG (an INFO field), which should be ignored by researchers.
- Remediation (ETA 2022):
  - We will remove AS_YNG in future releases.

6. WGS variant calls on chromosome Y need additional filtering

- Affects:
  - WGS joint callset VCFs
  - WGS joint callset Hail MT
- Suggested Action:
  - If you do not use variant calls on chrY, then no action.
  - Otherwise, we recommend that you use AD, GQ, and GT to filter variants on chromosome Y.
- Description: We see variants with heterozygous calls in chromosome Y, which cannot be correct germline calls. After manual review, we believe that regions of chromosome Y are prone to misalignment artifacts (low mappability). This will cause heterozygous calls in chrY that are likely artifacts. We have not investigated whether these are somatic mutations.
- Remediation (ETA 2023): We will provide a set of regions (via a BED file) that researchers can use to mask regions of the genome with poor calling accuracy for chromosome Y.

7. Small subset of samples missing corresponding CDR data

- Affects:
  - WGS joint callset VCFs
  - WGS joint callset Hail MT
  - Array single sample VCFs
  - Array merged Hail MT
  - Array PLINK bed/bim/fam
- Suggested Action:
  - If you are not using CDR data (e.g., surveys, EHR), then no action.
  - Otherwise, remove samples without corresponding CDR data. We will provide the lists of WGS and array samples without corresponding data in the CDR.
- Description: Due to an internal error in querying (since fixed) for the C2022Q2R2 CDR release, additional participants were dropped from the CDR that were not reflected in the genomic data. This affects 32 WGS (0.03%) samples and 55 array samples (0.03%).
  - Note: The affected participants are consented to appear in the genomic data.
- Remediation: We have fixed the source of this issue and this will not affect future releases. We have improved our processes in order to catch this type of issue earlier. We will provide two lists (WGS and array) of the affected samples through the User Support Hub.

FAQ

1. Why do you fail samples based on contamination rate for WGS, but not for array samples?

WGS analyses (e.g., mosaicism) rely on other signals, such as read counts, which are affected by contamination. Low rates of contamination do not affect array calls and problematic levels of contamination will be reflected in the array call rate.

2. Did you remove samples from participants with bone marrow transplants?

   Yes, we removed both array and WGS samples associated with participants that have received bone marrow transplants, according to the corresponding electronic health record (EHR) and survey responses provided by participants(Overall Health).

3. Are all samples in the WGS joint callset sourced from blood?

   Yes. Although the program does have saliva WGS samples, we did not include these samples in the Q2 2022 release. Once we identify any batch effects between saliva and blood samples (ETA 2022), we will reassess the inclusion of saliva samples in the joint WGS callset. If we decide that the batch effects will have minimal impact, we will include saliva samples in the WGS joint callsets in 2023.

References

[1] ***All Of Us User Support Hub*** (access required) https://aousupporthelp.zendesk.com/hc/en-us

[2] E Venner, D Muzny, et al., ***Whole-genome sequencing as an investigational device for return of hereditary disease risk and pharmacogenomic results as part of the All of Us Research Program***, *Genome Medicine* (2022). https://doi.org/10.1186/s13073-022-01031-z

[3] ***Detecting sample swaps with Picard tools – GATK.*** (n.d.). Retrieved October 21, 2021, from https://gatk.broadinstitute.org/hc/en-us/articles/360041696232-Detecting-sample-swaps-with-Picard-tools

[4] Pedersen and Quinlan, **Who's Who? Detecting and Resolving Sample Anomalies in Human DNA Sequencing Studies with Peddy** The American Journal of Human Genetics (2017) http://dx.doi.org/10.1016/j.ajhg.2017.01.017

[5] ***Illumina GenCall Data Analysis Software.*** (n.d.). Retrieved October 21, 2021, from https://www.illumina.com/Documents/products/technotes/technote_gencall_data_analysis_software.pdf.

[6] **CollectArraysVariantCallingMetrics (Picard)**, Retrieved October 21, 2021 , from https://gatk.broadinstitute.org/hc/en-us/articles/360037593871-CollectArraysVariantCallingMetrics-Picard-

[7] Zhang F, et al. **Ancestry-agnostic estimation of DNA sample contamination from sequence reads**. *Genome Research* (2020). https://doi.org/10.1101/gr.246934.118

[8] G. Jun et al., ***Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data***, American journal of human genetics doi:10.1016/j.ajhg.2012.09.004 (volume 91 issue 5 pp.839 - 848)

[9] ***Phred-scaled quality scores – GATK.*** (n.d.). Retrieved January 31, 2022, from https://gatk.broadinstitute.org/hc/en-us/articles/360035531872-Phred-scaled-quality-scores.

[10] Van der Auwera GA & O'Connor BD. (2020). ***Genomics in the Cloud: Using Docker, GATK, and WDL in Terra (1st Edition)***. O'Reilly Media. P.400

[11] ***gnomAD v3.1 New Content, Methods, Annotations, and Data ....*** (n.d.). Retrieved February 1, 2022, from https://gnomad.broadinstitute.org/news/2020-10-gnomad-v3-1-new-content-methods-annotations-and-data-availability.

[12] Van der Auwera GA & O'Connor BD. (2020). ***Genomics in the Cloud: Using Docker, GATK, and WDL in Terra (1st Edition)***. O'Reilly Media. P.166

[13] ***Relatedness - Hail.*** (n.d.). Retrieved October 21, 2021, from https://hail.is/docs/0.2/methods/relatedness.html#hail.methods.pc_relate.

[14] ***Which training sets arguments should I use for running VQSR ....*** (n.d.). Retrieved February 1, 2022, from https://gatk.broadinstitute.org/hc/en-us/articles/4402736812443-Which-training-sets-arguments-should-I-use-for-running-VQSR-.

[15] ***Resource bundle – GATK.*** (n.d.). Retrieved February 1, 2022, from https://gatk.broadinstitute.org/hc/en-us/articles/360035890811-Resource-bundle.

[16] ***The Omni Family of Microarrays.*** (n.d.). Retrieved February 16, 2022, from https://www.illumina.com/Documents/products/datasheets/datasheet_gwas_roadmap.pdf.

[17] International HapMap Consortium. **The International HapMap Project**. Nature. 2003 Dec 18;426(6968):789-96. doi: 10.1038/nature02168. PMID: 14685227.

[18] The 1000 Genomes Project Consortium, ***A global reference for human genetic variation***, Nature 526, 68-74 (01 October 2015) doi:10.1038/nature15393

[19] Mills R.E. et al. ***An initial map of insertion and deletion (INDEL) variation in the human genome***. Genome Res. 2006;16:1182–1190. doi:10.1101/gr.4565806.

[20] Krusche, P., Trigg, L., Boutros, P.C. *et al.* ***Best practices for benchmarking germline small-variant calls in human genomes***. *Nat Biotechnol* **37,** 555–560 (2019). https://doi.org/10.1038/s41587-019-0054-x

[21] Karczewski, K.J., Francioli, L.C., Tiao, G. *et al.* **The mutational constraint spectrum quantified from variation in 141,456 humans**. *Nature* 581**,** 434–443 (2020). https://doi.org/10.1038/s41586-020-2308-7

[22] M'Charek, A. ***The Human Genome Diversity Project: An Ethnography of Scientific Practice*** (Cambridge Studies in Society and the Life Sciences). Cambridge: Cambridge University Press. (2005) doi:10.1017/CBO9780511489167

[23] Ho, TK . ***Random Decision Forests***. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282.

[24] Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, *Journal of Machine Learning Research* 12, pp. 2825-2830, (2011).

[25] ***Downloads | gnomAD.*** (n.d.). Retrieved February 1, 2021, from https://gnomad.broadinstitute.org/downloads#v3-hgdp-1kg.

[26] ***Genetics - Hail.*** (n.d.). Retrieved October 21, 2021, from https://hail.is/docs/0.2/methods/genetics.html#hail.methods.hwe_normalized_pca.
[27] Erdős, P. **On cliques in graphs**, Israel Journal of Mathematics, 4 (4): 233–234, (1966), doi:10.1007/BF02771637, MR 0205874, S2CID 121993028
[28] Green RC, Berg JS, Grody WW, Kalia SS, Korf BR, Martin CL, et al. ***ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing.*** Genet Med. 15:565–574. (2013)
[29] ***"Picard Tools - By Broad Institute."*** (n.d.). Retrieved October 21, 2021 https://broadinstitute.github.io/picard
[30] Laurie CC, Doheny KF, et al. ***Quality control and quality assurance in genotypic data for genome-wide association studies***. Genet Epidemiol. 2010 Sep;34(6):591-602. doi: 10.1002/gepi.20516. PMID: 20718045; PMCID: PMC3061487.

Appendix A: Ancestry

We computed categorical ancestry for all of the WGS samples in *All of Us* and made these available to researchers. These predictions are also the basis for population allele frequency calculations in the Variant Annotation Table (e.g. gvs_afr_ac) and data in the Genomic Variants section of the public Data Browser. We used the high-quality set of sites (HQ sites), described in Appendix B, to determine an ancestry label for each sample. The ancestry categories are based on the same labels used in gnomAD [21], Human Genome Diversity Project [22], and 1000 Genomes [18]:

- African (afr)
- Latino/Native American/Ad Mixed American (amr)
- East Asian (eas)
- Middle Eastern (mid)
- European (eur) -- Composed of Finnish (FIN) and Non-Finnish European (NFE)
- Other (oth) -- not belonging to one of the other ancestries or is an admixture.
- South Asian (sas)

We trained a random forest classifier [23,24] on a training set of the HGDP and 1kg samples variants on chromosomes 20 and 21, obtained from gnomAD [25]. We generated the first 16 principal components (PCs) of the training sample genotypes (using the hwe_normalized_pca in Hail [26]) at the high-quality variant sites (see Appendix B) for use as the feature vector for each training sample. We used the truth labels from the sample metadata, which can be found alongside the VCFs. Note that we do not train the classifier on the samples labeled as "Other." We use the label probabilities ("confidence") of the classifier on the other ancestries to determine ancestry of "Other".

To determine the ancestry of *All of Us* samples, we project the *All of Us* samples into the PCA space of the training data and apply the classifier (see Figure A.1). Since we do not have truth labels, we can not determine the accuracy of our *All of Us* predictions. As a proxy for the accuracy of our *All of Us* predictions, we look at the concordance between the survey results and the predicted ancestry. The ancestry predictions can be found in Table A.1.

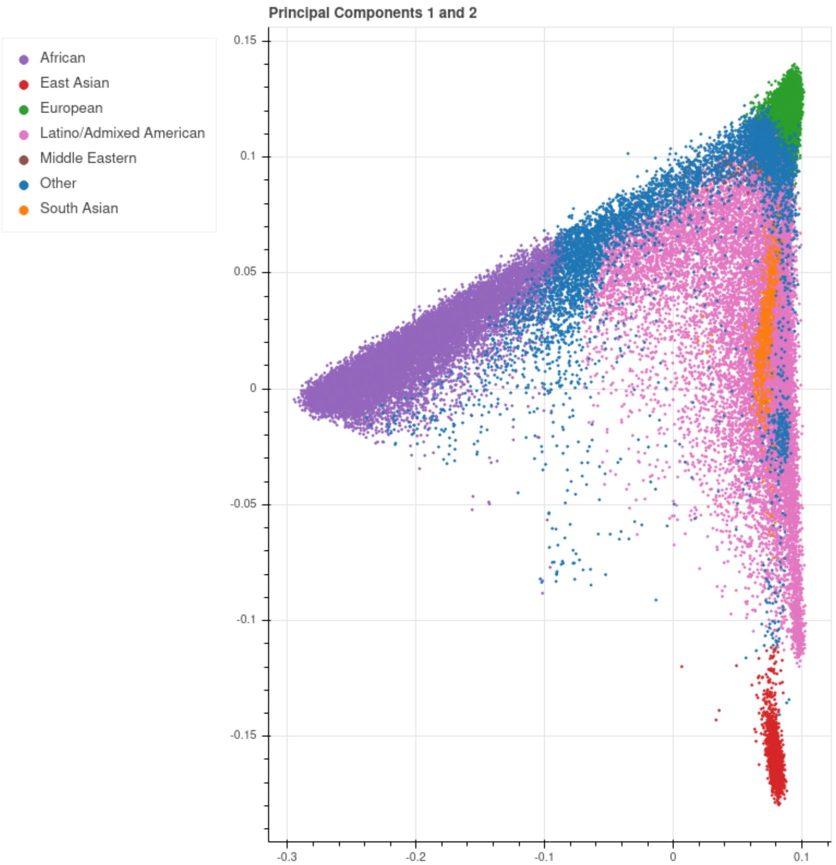


Figure A.1 -- Ancestry predictions for the *All of Us* WGS samples plotted on the first two principal components (PC1 on x-axis and PC2 on the y-axis) of the genotype calls.

Table A.1 -- Breakdown of the computed ancestries in *All Of Us* WGS data

Computed Ancestry (sorted by percentage)	Count (percentage)
European	48112 (48.8%)
African	23179 (23.5%)
Latino/Admixed American	15133 (15.3%)
Other	8907 (9.0%)
East Asian	2119 (2.1%)
South Asian	973 (1.0%)
Middle Eastern	167 (0.2%)

Total: 98590 (100.0%)

We evaluated the performance of the ancestry predictions using two different test datasets:

1. A holdout set of training samples. We tested performance with and without the “Other” ancestry
    a. Error rate (incl Other): 0.046
        i. See Table A.2
        ii. Please see Known Issue #3, since the error rate is higher for Middle Eastern (mid) ancestry. Our classifier conflates Middle Eastern and Other.
    b. Error rate (not incl Other): 0.009
        i. See Table A.3

Table A.2 -- Error rate (incl. Other) on labeled training data using holdout set

	Predicted						
Actual	AFR	AMR	EAS	EUR	MID	OTH	SAS
AFR	200	0	0	0	0	0	0
AMR	0	50	0	0	0	0	0
EAS	0	0	198	0	0	2	0
EUR	0	0	0	199	0	1	0
MID	0	0	0	0	34	16	0
OTH	1	0	2	3	10	25	6
SAS	0	0	0	0	0	3	197

Table A.3 -- Error rate (not incl. Other) on labeled training data using holdout set

	Predicted					
Actual	AFR	AMR	EAS	EUR	MID	SAS
AFR	200	0	0	0	0	0
AMR	0	50	0	0	0	0
EAS	0	0	200	0	0	0
EUR	0	0	0	199	0	0
MID	0	0	0	6	44	0
SAS	0	1	1	0	0	199

2. We evaluated the performance of the ancestry predictions against the self-reported ethnicity of the *All of Us* samples as ground truth. The performance should be worse

than the holdout HGDP samples, but this is expected. Self-reported ethnicity does not correspond to the populations listed above and is prone to false reporting.

“Correct” labeling between HGDP/1kg populations and *All of Us* ethnicities:

1. African (AFR) → Black
2. Latino/Ad Mixed American (AMR) → Hispanic
3. East Asian (EAS) → Asian
4. Finnish (FIN) → White
5. Middle Eastern (MID) → MENA
6. Non-Finnish European (NFE) → White
7. Other (OTH) → Other (do not include skipped)
8. South Asian (SAS) → Asian

We do not include any samples where the self-reported ethnicity is “Skip”, “Prefer not to answer”, or was not filled in. If a participant selected that their ethnicity was not a possible selection (“NoneOfThese”), we counted them as “Other”.

Based on the procedure above, the concordance between self-reported ethnicity and the ancestry predictions: 0.877

Appendix B: High quality site determination

In order to do relatedness and ancestry checks, we identified a corpus of sites that can be called accurately in both our ancestry training set (HGDP+1KG) and our target data (*All of Us* WGS callset). We used a similar methodology that gnomAD used to determine high-quality sites [11], but we repeat it here for clarity:

1. Autosomal, bi-allelic single nucleotide variants (SNVs) only
2. Allele frequency > 0.1%
3. Call rate > 99%
4. LD-pruned with a cutoff of r2 = 0.1

Our aim was to assemble a set of independent sites where we can be confident of the accuracy.

We identified 56695 high-quality (HQ) sites in the Q2 2022 callset. These were HQ sites in both the HGDP+1kg training VCF and the *All of Us* Q2 2022 callset. A sites-only VCF of the HQ sites is available in the RWB (access required).

Appendix C: Relatedness

We calculated the kinship score and reported any pairs with a kinship score over 0.1.
The kinship score is half of the fraction of the genetic material shared (ranges from 0.0 - 0.5).

- Parent-child or siblings: 0.25
- Identical twins: 0.5

Please see the Hail pc_relate function [13] documentation for more information, including interpretation.
We will determine the maximal independent set [27] for related samples to minimize the number of samples that would need pruning. Using the HQ sites identified in Appendix B, researchers can remove first and second degree relatives.

We estimated 4,846 related pairs and 4,069 samples in the maximal independent set for kinship scores above 0.1. The sample pairs, with kinship score, and the set are available in the RWB (access required).

Appendix D: Samples used in the Sensitivity and Precision Evaluation

In order to calculate the sensitivity and precision of the joint callset, we included four well-characterized samples in the Q2 2022 callset (Table D.1). We sequenced the NIST reference materials (DNA samples) from Genome in a Bottle (GiaB) and performed variant calling as described in the main text. We used the corresponding published set of variant calls for each sample as the ground truth in our sensitivity and precision calculations [20].

Please note that the control samples do not appear in the data released to researchers.

Table D.1 -- Samples used in sensitivity and precision evaluation

Control Sample	Ground Truth	Genome Center	GVCF origin	Notes
HG-001	GiaB	BI	DRAGEN 3.4.12	NA12878
HG-003	GiaB	UW	DRAGEN 3.4.12	Ashkenazi Trio NA24149 - Father
HG-004	GiaB	BI	DRAGEN 3.4.12	Ashkenazi Trio NA24143 - Mother
HG-005	GiaB	BI	DRAGEN 3.4.12	Han ancestry NA24631- Son

Genome Center:
BI -- Broad Institute
UW -- University of Washington

Appendix E: Single sample QC processes performed

See Table E.1 to determine which of the single sample QC processes were performed. In cases where both GCs and the DRC performed a check, if the sample failed either check, it was not included in the Q2 2022 release (though see Known Issues above for exceptions regarding call rate).

Table E.1 -- Single sample QC processes and which centers performed the check

QC process	Data types	Calculated at the DRC or GCs?
Fingerprint Concordance	WGS	Both*
Sex concordance	Arrays	GCs only
Sex concordance	WGS	Both
Cross-individual contamination rate	Arrays	GCs only
Cross-individual contamination rate	WGS	Both
Call rate	Arrays	GCs only
Coverage	WGS	GCs only

*One GC (Broad Institute) performed an internal check against a different fingerprint (Fluidigm SNP genotyping (SNPtype chemistry) using the 96.96 Dynamic Array), which did not use the same fingerprint sites as the array. The DRC treated these samples the same as from the other GCs and ran the array concordance as described in the main text of this document.

Appendix F: *All of Us* Hereditary Disease Risk genes

The following gene symbols are in the *All of Us* Hereditary Disease Risk (AoUHDR) genes. We have additional WGS QC criteria in the regions covered by these genes, described in Table 1 of the main text. In the Q2 2022 callset, the AoUHDR genes are the same as the American College of Medical Genetics and Genomics' list of 59 genes where incidental findings should be reported (ACMG59) [28]. The AoUHDR gene list may change in future releases.

ACTA2, ACTC1, APC, APOB, ATP7B, BMPR1A, BRCA1, BRCA2, CACNA1S, COL3A1, DSC2, DSG2, DSP, FBN1, GLA, KCNH2, KCNQ1, LDLR, LMNA, MEN1, MLH1, MSH2, MSH6, MUTYH, MYBPC3, MYH11, MYH7, MYL2, MYL3, NF2, OTC, PCSK9, PKP2, PMS2, PRKAG2, PTEN, RB1, RET, RYR1, RYR2, SCN5A, SDHAF2, SDHB, SDHC, SDHD, SMAD3, SMAD4, STK11, TGFBR1, TGFBR2, TMEM43, TNNI3, TNNT2, TP53, TPM1, TSC1, TSC2, VHL, and WT1

Appendix G: DRAGEN invocation parameters

Table G.1 summarizes the parameters used by the GCs to generate GVCFs, contamination estimates, and sex ploidy calls from the DRAGEN.

Table G.1 DRAGEN 3.4.12 parameters run at all GCs

Parameter	Parameter Value	Description

-f	n/a	Overwrite if output exists
-r	<hg38-ref-dir>	The reference to use
--fastq-list	<path-to>/fastq_list.csv	A list of fastq files to use as input for this sample
--fastq-list-sample-id	<sampleID>	The sample ID to use for naming this sample
--output-directory	<output-dir>	The location of the final output files
--intermediate-results-dir	<int-results-dir>	The location to write intermediate outputs
--output-file-prefix	[CenterID]_[Biobankid_Sampleid]_[LocalID:optional]_[Rev#]	Standardized naming prefix for each output file
--enable-variant-caller	TRUE	Turn on variant call outputs
--enable-duplicate-marking	TRUE	Mark duplicate reads during alignment
--enable-map-align	TRUE	Produce an alignment from unaligned read input
--enable-map-align-output	TRUE	Store the output of the alignment
--output-format	CRAM	Store the alignment as a CRAM file
--vc-hard-filter	DRAGENHardQUAL:all:QUAL<5.0;LowDepth:all:DP<=1'	This parameter setting changes the threshold on the quality to 5.
--vc-frd-max-effective-depth	40	Setting this parameter puts a limit on the penalty value that is applied for variant calls that deviate from the expected 50% allele fraction for heterozygous variants.
--qc-cross-cont-vcf	<path-to/SNP_NCBI_GRCh38.vcf>	Marker sites to use for contamination estimation
--qc-coverage-region-1	<path-to/wgs_coverage_regions.bed>	Regions to use for coverage analysis (whole genome)
--qc-coverage-reports-1	cov_report	The type of reports requested for qc-coverage-region-1
--qc-coverage-region-2	<path-to/HDRR_regions.bed>	Regions to use for coverage analysis (HDR reportable regions)
--qc-coverage-reports-2	cov_report	The type of reports requested for qc-coverage-region-2
--qc-coverage-region-3	<path-to/PGx_regions.bed>	Regions to use for coverage analysis (PGx reportable regions)
--qc-coverage-reports-3	cov_report	The type of reports requested for qc-coverage-region-3

Appendix H: Self-reported sex at birth

See Table H.1 for the counts and percentages of participant responses to the sex assigned at birth question in the Basics survey (based on *All of Us* CDR release C2022Q2R2). The survey question presented to participants was “What was your biological sex assigned at birth?” and can be found in the Basics survey. The CDR code for this question is sex_at_birth.

Table H.1 -- Q2 2022 release participants response breakdown to sex assigned at birth question

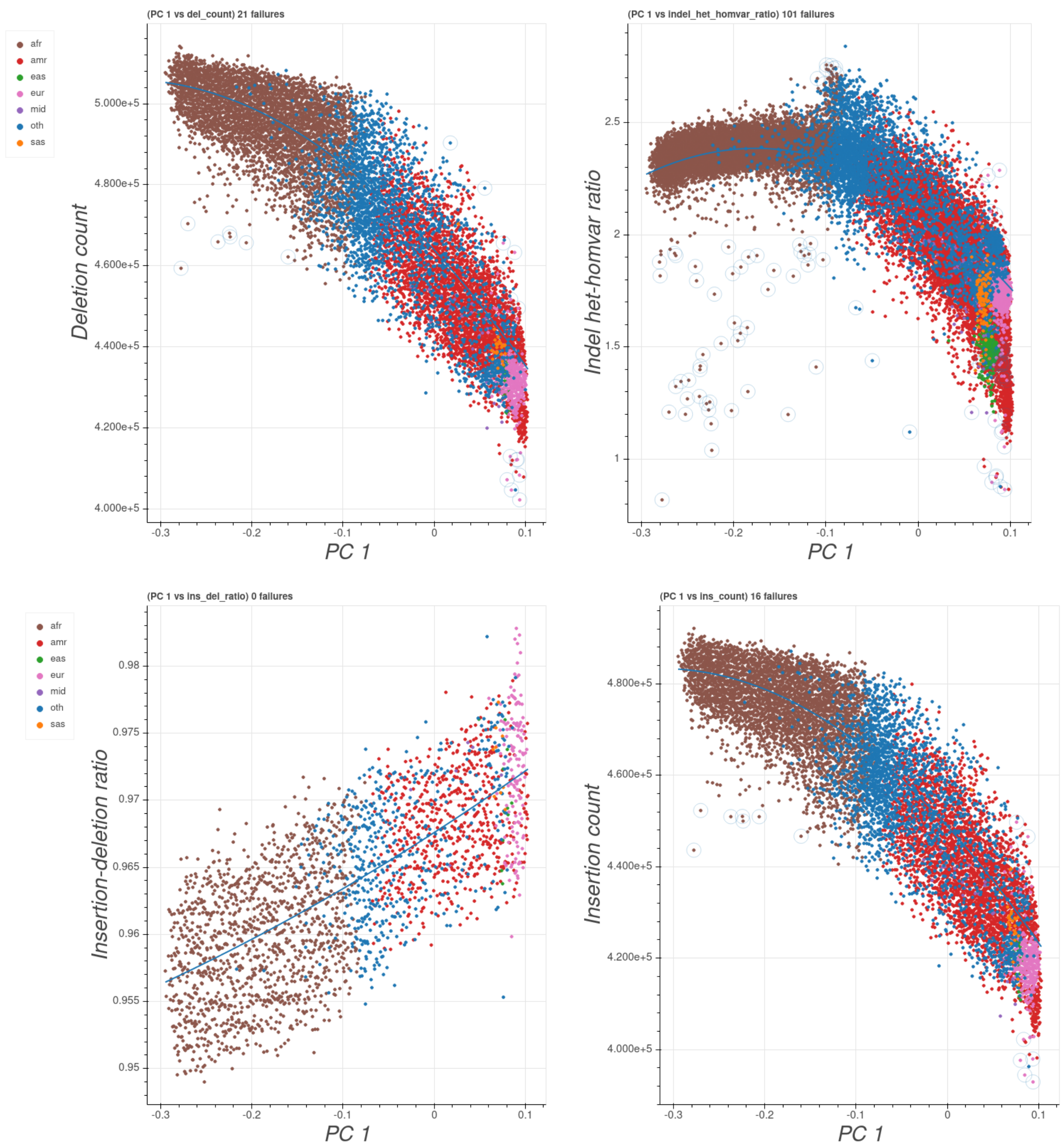
Q2 2022 Release	WGS		Array	
Sex assigned at birth responses	counts	percent	counts	percent
Female	58925	59.79	99346	60.18
Male	38133	38.69	63251	38.32
I prefer not to answer	79	0.08	120	0.07
None of these fully describe me	35	0.04	55	0.03
Intersex	22	0.02	35	0.02
No matching concept*	363	0.37	669	0.41
PMI: Skip*	1001	1.02	1596	0.97
Total	98558		165072	

Percentages may not add to 100 due to rounding. The total counts reflect the missing CDR samples in the Q2 2022 release (see Known Issue #7).

* “No matching concept” and “PMI: Skip” are separate counts both referring to no response for sex_at_birth. These are separate because participants in “No matching concept” did select a gender option for this survey question. The terms used here are the Concept Names as they appear in the CDR.

Appendix I: Plots of the first principal component against population outlier QC metrics

Figure I.1 (next page) contains the plots of the first principal component against metrics used for determining sample population outliers. Note that we use sixteen principal components for determining which samples should be flagged for being outliers in a metric. The blue line shows the linear regression fit in the first dimension (residuals are calculated as the distance from this hyperplane). The failure count over these plots will sum higher than the 156 flagged samples, since samples can get flagged for multiple criteria. Please see the next page for Figure I.1.



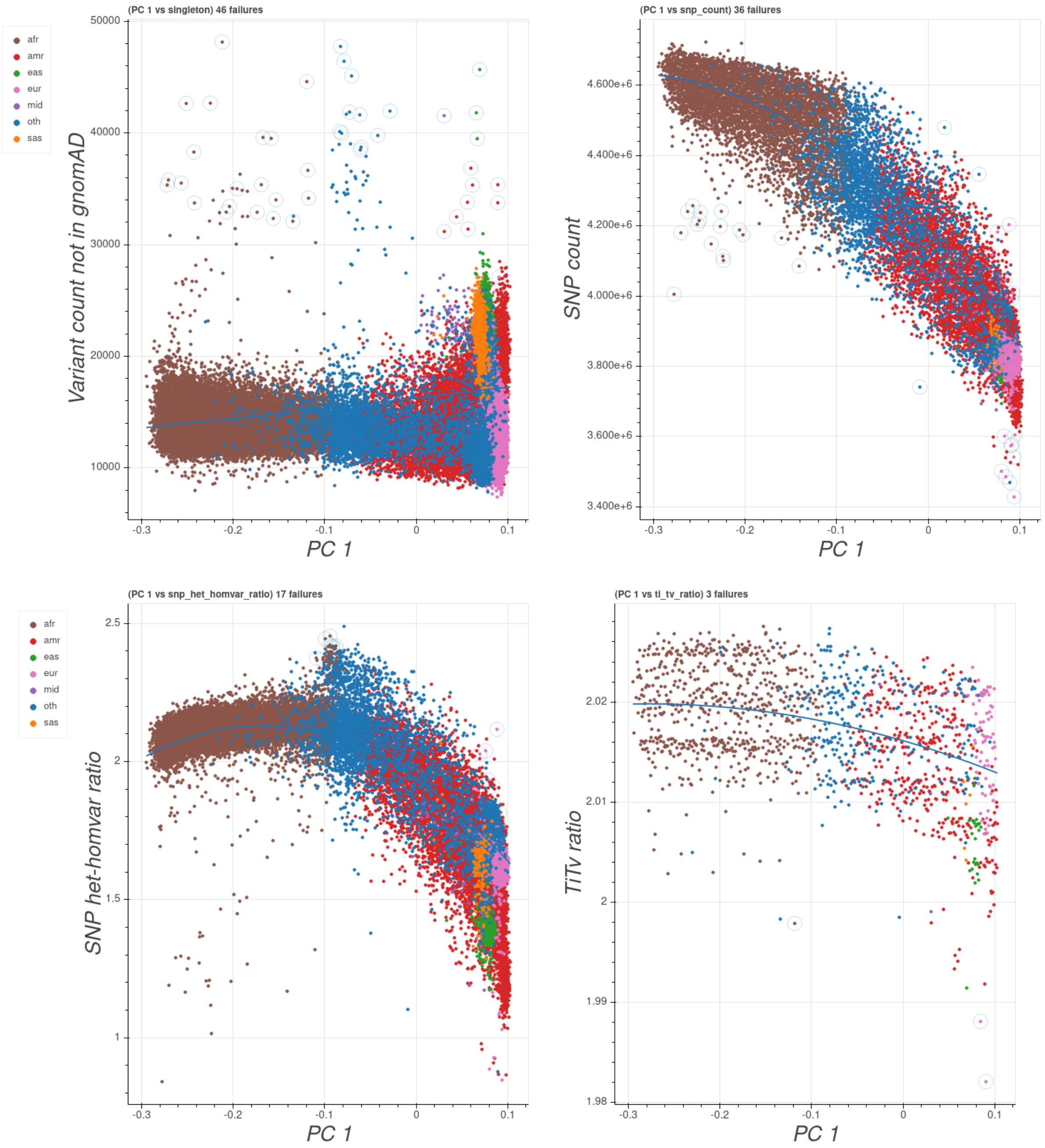


Figure I.1 -- Sample population outlier plots for eight metrics (see Population Outlier Flagging). Each metric (y-axis) is plotted against the first (of sixteen) principal components (x-axis). Outliers are identified by regressing out the principal components and determining if the residual is over 8 MADs from the sample population.

Appendix J: Array processing overview

See Figure J.1 for an overview of the array genotyping process for the *All of Us* Research Program. The three GCs used identical array products, scanners, resource files, and genotype calling software. The GCs used the Illumina Global Diversity Array (GDA) (https://www.illumina.com/products/by-type/microarray-kits/infinium-global-diversity.html).

- Array product details:
  - Bead pool file: GDA-8v1-0_A5.bpm
  - EGT cluster file: GDA-8v1-0_A1_ClusterFile.egt
  - gentrain v.3
  - reference hg19 (Note: We liftover to hg38 before publishing array data in the RWB)
  - gencall cut-off 0.15
  - 1,914,935 assays
    - 44,172 indels
    - 9,935 IntensityOnly (probes intended only for Copy Number Variant (CNV) calling)
    - 70,174 duplicates (same position, different probes)
- Chemistry: Illumina Infinium LCG using automated protocol
- Liquid handling robotics: Various platforms across the genome centers
- Scanners: Illumina iSCANs with Automated Array Loader
- Software:
  - Illumina IAAP Version: iaap-cli-linux-x64-1.1.0-sha.80d7e5b3d9c1fdfc2e99b472a90652fd3848bbc7.tar.gz
    - IAAP converts raw data (.idat files – 2 per sample) into a single .gtc file per sample using the .bpm file (defines strand, probes sequences, and illumicode address) and the .egt file (defines the relationship between intensities and genotype calls)
  - Picard-2.20.X or above [29], but exact version depended on the GC.
    - Johns Hopkins: 2.20.8-SNAPSHOT
    - Broad Institute: 2.23.0
    - University of Washington: 2.23.3
    - Picard versions 2.23.0 and above modified GtcToVcf to read the gtc_call_rate from the GTC file and put it into the VCF header.
      - Please see Known Issue #1 and Known Issue #2 for issues that arose due to Picard version inconsistencies.
    - Picard tool, GTCtoVCF, converts the .gtc file into a vcf file.
  - BAFRegress version 0.9.3 [8]
    - BAFRegress measures the within species DNA sample contamination using B allele frequency data from Illumina genotyping arrays using a regression model
- Quality Control: Each genome center ran the GDA array under Clinical Laboratory Improvement Amendments (CLIA) compliant protocols. We generated .gtc files and uploaded metrics to in-house Laboratory Information Management Systems (LIMS) systems for quality control review. At batch level (each set of 96 well plates run together in the laboratory at one time), each GC included positive control samples, which were required to have > 98% call rate and >99% concordance to existing data, in order to

approve release of the batch of data. At the sample level, the call rate and sex are the key quality control determinants [30]. Contamination is also measured using BAFRegress [8] and reported out as metadata. Any sample with a call rate below 98% is repeated one time in the laboratory. Genotyped sex is determined by plotting normalized X versus normalized Y intensity values for a batch of samples [30]. Any sample discordant with 'sex assigned at birth' reported by an *All of Us* participant is flagged for further detailed review. If multiple sex discordant samples are clustered on an array or on a 96 well plate, the entire array or plate will have data production repeated. Samples identified with sex chromosome aneuploidies are also reported back as metadata (XXX, XXY, XYY, etc). A final processing status of "PASS," "FAIL" or "ABANDON" is determined before release of data to the DRC. An array sample will PASS if the call rate is > 98% and the genotyped sex and sex assigned at birth are concordant (or the sex assigned at birth is "Intersex", "I prefer not to answer", "none of these fully describe me", or skipped the question). An array sample will FAIL if the genotyped sex and the sex assigned at birth are discordant or if the call rate is less than 98% on the first run of the sample. An array sample will have the status ABANDON if the call rate is less than 98% after at least 2 attempts at the GC.

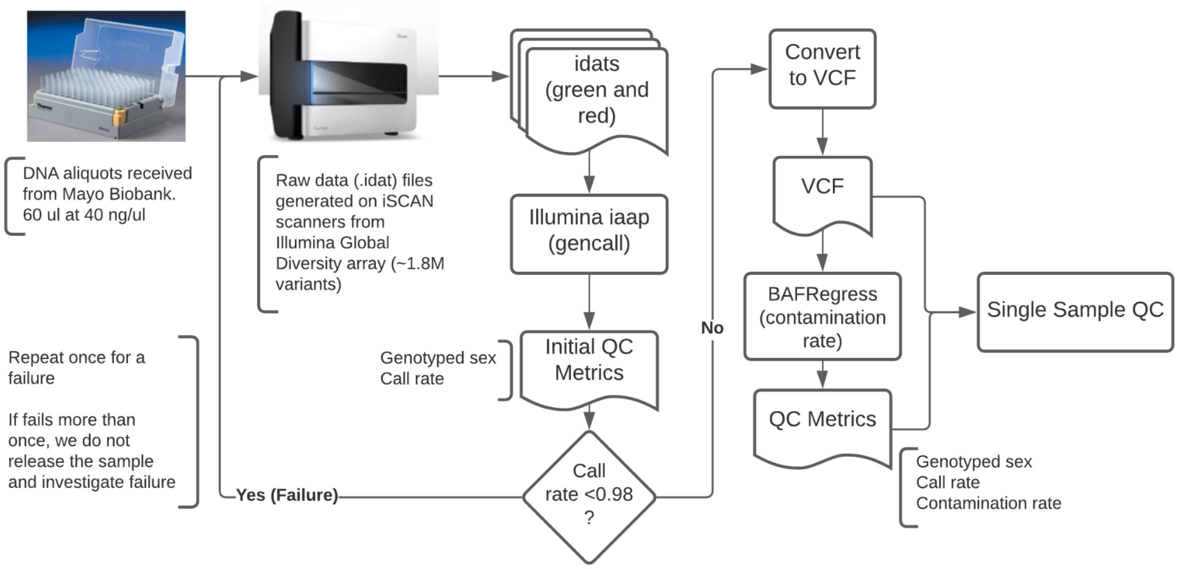

DNA aliquots received from Mayo Biobank. 60 ul at 40 ng/ul
Raw data (.idat) files generated on iSCAN scanners from Illumina Global Diversity array (~1.8M variants)
idats (green and red)
Illumina iaap (gencall)
Genotyped sex Call rate
Initial QC Metrics
Call rate <0.98 ?
Yes (Failure)
No
Repeat once for a failure
If fails more than once, we do not release the sample and investigate failure
Convert to VCF
VCF
BAFRegress (contamination rate)
QC Metrics
Single Sample QC
Genotyped sex Call rate Contamination rate

Figure J.1 -- Overview of the array processing pipeline.

Appendix K: Self-reported race/ethnicity

As seen in Table K.1, the race/ethnicity breakdown of the genomic data is similar to all participants in the *AoURP* (based on *All of Us* CDR release C2022Q2R2). Samples with "PMI_Skip" responses include participants that answered "prefer not to answer", entered blank text, or did not respond to the survey question. As seen in Appendix L, all WGS and array samples have corresponding survey data.

Table K.1 -- Self-reported Race/Ethnicity breakdown of the WGS samples

Self-Reported Race/Ethnicity	Survey Response Counts (%)	WGS Counts (%)	Array Counts (%)
AIAN	47 (0.0%)	–	–
Asian	12317 (3.3%)	2980 (3.0%)	5163 (3.1%)
Asian, White	858 (0.2%)	213 (0.4%)	359 (0.4%)
Black	72706 (19.5%)	21322 (21.6%)	32347 (19.6%)
Black, White	1064 (0.3%)	283 (0.6%)	443 (0.5%)
Hispanic	59085 (15.9%)	17325 (17.6%)	26206 (15.9%)
Hispanic, White	2851 (0.8%)	677 (1.4%)	1139 (1.4%)
MENA	2122 (0.6%)	522 (0.5%)	907 (0.6%)
Other	14606 (3.9%)	3502 (2.4%)	5816 (2.4%)
PMI_Skip	6732 (1.8%)	1885 (1.9%)	3004 (1.8%)
White	200018 (53.7%)	49849 (50.6%)	89688 (54.3%)
Total	**331382 (100.0%)**	**98558 (100.0%)**	**165072 (100.0%)**

Appendix L: Data type availability with genomic data

We provide 95,596 WGS samples (97%) with corresponding array data (see Known Issues #1 for why this is not 100% of WGS samples). Additionally, both WGS (Table L.1) and array (Table L.2) data have other corresponding non-genomic data. This can be one or more of the following:

- Electronic Health Records (EHR)
- Physical Measurements (PM)
- Participant Provided Information (PPI/surveys)
- Fitbit (FB)

Descriptions of the non-genomic data can be found on the *All of Us* Data Sources page.

Table L.1 -- WGS overlap with non-genomic data types

Data Combination	Description	Participant Count
WGS	any WGS data	98590

WGS and PPI	any WGS AND any PPI	98558
WGS and PPI and PM	any WGS AND any PPI AND any PM	98463
WGS and EHR	any WGS AND any EHR	81054
WGS and PPI and EHR	any WGS AND any PPI AND any EHR	81054
WGS and PPI and EHR and PM	any WGS AND any EHR AND any PM AND any PPI	81023
WGS and Fitbit	any WGS AND any Fitbit	3378
WGS and PPI and Fitbit	any WGS AND any PPI AND Fitbit	3378
WGS and PPI and PM and Fitbit	any WGS AND any PPI AND any PM AND any Fitbit	3373
WGS and Fitbit and PPI and EHR	any WGS AND any Fitbit AND and PPI AND any EHR	2817
WGS and PPI and EHR and PM and Fitbit	any WGS AND any EHR AND and PM AND any PPI AND any Fitbit	2817

Table L.2 -- Array overlap with non-genomic data types

Data Combination	Description	Participant Count
Array	any Array data	165127
Array and PPI	any Array AND any PPI	165072
Array and PPI and PM	any Array AND any PPI AND any PM	164608
Array and EHR	any Array AND any EHR	136922
Array and PPI and EHR	any Array AND any PPI AND any EHR	136922
Array and PPI and EHR and PM	any Array AND any EHR AND any PM AND any PPI	136877
Array and Fitbit	any Array AND any Fitbit	6810
Array and PPI and Fitbit	any Array AND any PPI AND Fitbit	6810
Array and PPI and PM and Fitbit	any Array AND any PPI AND any PM AND any Fitbit	6779
Array and Fitbit and PPI and EHR	any Array AND any Fitbit AND and PPI AND any EHR	5678
Array and PPI and EHR and PM and Fitbit	any Array AND any EHR AND and PM AND any PPI AND any Fitbit	5677

Appendix M: Genome Centers and Data and Research Center

Below is the listing of the three Genome Centers (GCs), the Data and Research Center (DRC), and the Biobank.

Role	Institution(s)	PI(s)
Genome Center	Baylor College of Medicine, Johns Hopkins University	Richard Gibbs Eric A. Boerwinkle Kimberly F. Doheny
	Broad Institute	Stacey Gabriel
	Northwest Genomics Center at the University of Washington	Deborah A. Nickerson
Data and Research Center	Vanderbilt University Medical Center	Paul Harris Dan M. Roden
	Broad Institute	Anthony Philippakis
	Verily Life Sciences	David Glazer
Biobank	Mayo Clinic	Stephen Norman Thibodeau