Imputation Services

The *All of Us* + AnVIL Imputation Services: The world's largest, most ancestrally diverse genomic reference panel for equitable variant discovery

SLABRECQ@BROADINSTITUTE.ORG | Sofia Labrecque, Morgan Aster, Jose Soto, Saloni Shah, Matt Bemis, Chris Kachulis, Sam Lee, Lee Lichtenstein

OVERVIEW

Genotyping arrays and low-pass WGS (lpWGS) (including blended genome-exome sequencing or BGE) cut per-sample sequencing costs by up to 30×, but leave variants unobserved. Imputation fills the gap — and its quality scales directly with the size, accuracy, and diversity of the reference panel.

The problem: existing services rely on small, predominantly European panels and slow on-premises infrastructure. Over 80% of GWAS participants have been of European ancestry, a chronic imbalance that limits rare variant discovery and PRS portability across populations.[1]

The *All of Us* + AnVIL Aray + Low-Pass Whole Genome Sequencing and Blended Genome-Exome Imputation Service addresses both. With 515,000+ whole genomes — more than 250,000 from non-European inferred ancestries — running on FedRAMP Moderate-compliant cloud infrastructure, it delivers superior imputation accuracy where it has historically mattered least.

Imputation of structural variants (SVs) - large insertions, deletions, and other rearrangements - comes with its own challenges. SVs are a major driver of disease risk and phenotypic variation but are difficult to call reliably from standard short-read sequencing. Imputing SVs from short-read Whole Genome Sequencing (WGS) data would make this class of variation broadly accessible without the cost of long-read sequencing at scale. The *All of Us* Structural Variant Imputation Service, available in November 2026, enables accurate imputation of structural variant data from a reference panel of 12,000+ long-read whole genome sequences, including imputation of some areas of the genome that are particularly difficult to call.

The *All of Us* + AnVIL Array and lpWGS/BGE Imputation Services reference panel has

- **>250K** Non-European inferred ancestries
- **>515K** Whole genomes
- **>990M** Variants

The *All of Us* SV Imputation Service reference panel has

- **>12k** *AoU* Long Read samples
- **~1.2M** SVs, ranging from 50bp to 10kbp
- **~125M** SNVs and indels (including SNVs and indels in regions hard to call with short-reads!)

HOW DOES IT WORK?

1) Submit multi-sample VCF (array), cram (lpWGS/BGE), or GVCF (SV) via GUI or CLI — cloud-hosted inputs supported
2) Automated QC catches malformed inputs and reference mismatches within 30 min — user is not charged
3) Without needing to be an *All of Us* researcher, Phase & impute against the protected *All of Us* + AnVIL panel (array and lpWGS/BGE) or the *All of Us* long-reads panel (SV) using RESHAPE (re-identification defense) inside Terra's FedRAMP environment — panel haplotypes are never exposed
4) Download results within ~2 days

Get started

Check out our Pre-print

WHAT IS THE PRICE?

Pipeline	Free Quota (NIH-funded; 1 sample = 1 quota)	Per Sample Price (Non-Profit / Commercial)
Array	2,500	$0.30 / $0.40
lpWGS/BGE	300	$3.50 / $4.00
SV	500	$1.50 / $2.00

1. Abul-Husn, N.S., and Kenny, E.E. (2019). Personalized Medicine and the Power of Electronic Health Records. Cell 177, 58–69.

WHY CHOOSE THE *All of Us* + AnVIL IMPUTATION SERVICES?

Imputation can help complete your datasets **efficiently** and **accurately.**

World's largest, most diverse dataset for superior accuracy
515K+ high-quality genomes delivering top-tier accuracy, especially in underrepresented ancestry groups and rarer variants.

Scalable service
The cloud-native service scales to match the size of your genomic research.

Security
FISMA and FedRAMP Moderate. Industry-level infosec best practices.

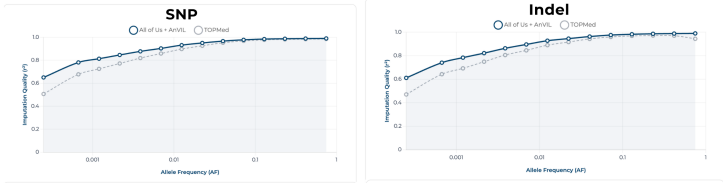
Fast turnaround
Turnaround time within days (vs. weeks to months using other services.)

Interface for everyone
Both CLI for computational researchers and GUI for low-code experiences.

HAS THE IMPUTATION SERVICE BEEN SCIENTIFICALLY VALIDATED?

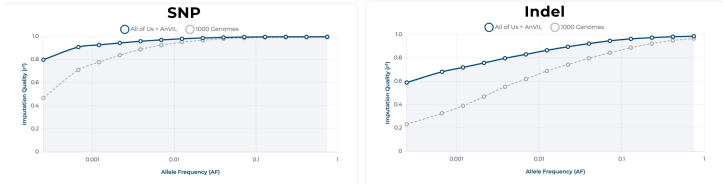
Array data

Aggregate r^2 for SNPs and INDELs across 42 held-out samples, from diverse reference populations, benchmarked against TOPMed



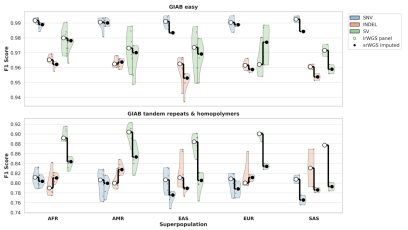
lpWGS/BGE data

Aggregate r^2 for SNPs and INDELs across 42 held-out samples, from diverse reference populations, benchmarked against 1000 Genomes



Structural variant imputation

F1 scores against assembly-based truth for SNVs, indels, and SVs on chromosome 20, in both the long-read panel and short-read held-out imputation, across 25 HPRC2/HGSVC3 samples stratified by superpopulation



WHAT TYPES OF IMPUTATION ARE AVAILABLE?

NOW	NOW	SOON
Array data	Low-Pass WGS data	Structural Variants

ACKNOWLEDGEMENTS

This work was supported by funding from the National Institutes of Health and in collaboration with the National Human Genome Reseach Institute. The imputation panel generation and service were supported by awards from the *All of Us* (OT2OD032121, OT2OD002748, OT2OD003811, and OT2OD002750) and the National Human Genome Research Institute (SU2AHG010262 and SU2AHG010263). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We thank the TOPMed imputation service team and the AnVIL team for helpful discussion and suggestions. The Centers for Common Disease Genomics (CCDG) are funded by the National Human Genome Research Institute (UM1HG008853, UM1HG008895, UM1HG008898, UM1HG008901, and U24HG008956) and in part by the National Heart, Lung, and Blood Institute. We thank the participants and investigators of CCDG for their contributions to the imputation reference panel. Complete and detailed acknowledgments of all CCDG cohorts are listed in the Supplementary Information. We gratefully acknowledge *All of Us* participants for their contributions, without whom this research would not have been possible. We also thank the National Institutes of Health's *All of Us* Research Program for making available the participant data for these scientific tools.