Benchmarking and quality analyses on the *All of Us* short-read structural variant catalog

All of Us Curated Data Repository (CDR) release C2024Q3R3

Introduction

Widespread benchmarking of structural variants (SVs) from short-read whole genome sequencing (srWGS) remains a significant challenge for the field of human genetics. The *All of Us* cohort of samples with srWGS SV data is unique in the availability of matched genomic datasets (i.e. single-nucleotide variant [SNV] arrays and long-read genome sequencing [lrWGS]). There also exist a number of intrinsic measures that can be used to assess the technical quality of a dataset (e.g. comparisons to external datasets, assessments of inherited variation, comparison of allele frequency spectra and variant size distributions, and population genetic principles such as Hardy-Weinberg Equilibrium). Combining these methods, we assess seven measures of technical quality for the srWGS SV *All of Us* dataset, described in the Genomic Research Data Quality Report [1] and this supplemental document.

In the Genomic Research Data Quality Report [1], we describe:

1. Variant counts (cohort-wide and per-sample) relative to gnomAD V2 [2] and the most recent 1000 Genomes Project high-coverage srWGS callset [3]
2. Size distribution of SVs
3. Hardy-Weinberg equilibrium

In this supplemental benchmarking report, we additionally describe:

4. Linkage disequilibrium with srWGS SNVs and Indels
5. Patterns of evolutionary constraint
6. Benchmarking against lrWGS
7. Benchmarking against microarrays

Comparisons to SNVs and Indels

Linkage disequilibrium with SNVs and Indels

Data and Methods

Given that most common SVs segregate on haplotypes with distinct sets of SNV and small insertion and deletion variant (indel) calls, the presence of nearby SNVs and indels in linkage disequilibrium (LD) with SV calls is an indicator of SV callset quality.

To quantify this, we computed LD between the srWGS SV joint callset and SNVs and indels from the srWGS SNP and Indel joint callset. We conducted this analysis in Hail v0.2.130 in a Python notebook backed by a Spark 3.5.0 cluster with 12 non-preemptible and 24 preemptible worker nodes. Due to the high computational resources necessary for these calculations, LD analyses were conducted on a subset of 14,968 samples. We analyzed samples from the *All of Us* genetic ancestry groups with at least 900 samples in the SV callset: 1KGP-HGDP-EUR-like (EUR; n=47,884 samples in entire cohort), 1KGP-HGDP-AFR-like (AFR; n=24,222), 1KGP-HGDP-AMR-like (AMR; n=15,361), 1KGP-HGDP-EAS-like (EAS; n=2,054), and 1KGP-HGDP-SAS-like (SAS; n=914). Genetic ancestry group labels describe the genetic similarity of each group to a reference population based on the srWGS SNP and Indel data and are further described in the Genomic Research Data Quality Report Appendix G [1]. We randomly selected 4,000 samples from each genetic ancestry group that contained more than 4,000 samples in the cohort (AFR, AMR, and EUR) and selected all of the samples from the EAS and SAS ancestry groups to get a total subset of 14,968 samples. We analyzed LD between all SVs with PASS filter status and SNVs and indels with PASS filter status that had a minor allele frequency of at least 1% in either the full cohort or one of these genetic ancestry groups.

LD between the callsets was computed according to the same methods used for the *All of Us* CDRv7 benchmarking and quality analyses [4]. To recap, we computed LD by first constructing two matrices:

1. An m x n matrix A where m is the number of SV calls after minor allele frequency filtering, n is the number of samples in the cohort or genetic ancestry group, and A_{ij} is the number of alternate alleles for sample j at SV site i.
2. An s x n matrix B where s is the number of SNVs and indels after minor allele frequency filtering and B_{ij} is the number of alternate alleles for sample j at SNV/indel site i.

We defined LD as the R^2 of alternate allele dosage between each pair consisting of one SV site and one SNV site [5]. We calculated R^2 values by computing the matrix multiplication AB^T after mean-centering and variance-standardizing each matrix, and then squaring each entry of the resulting correlation matrix. We limited computation to SV/SNV pairs where the SNV was within 1 megabase of the SV by defining a window extending from 1 megabase (Mb) before the start position (POS) of the SV to 1 Mb after the end position (END). Then, correlations were computed between each SV and the SNVs located within the window using Hail's block matrix sparsification functionality. For each SV we identified the SNV with which the R^2 value was maximized. Given that previous LD analyses of SVs have shown that LD was much weaker for SVs that occurred in repetitive sequence contexts [2], we further subdivided the results according to the genomic context in which the SV occurs; we classified each SV as occurring in

segmental duplications (SD), simple repeats (SR), other repeat-masked sequence (RM), or the unique sequence (US) outside of RM using methods from Zhao et al. 2021 [6].

Results

A violin plot of the maximum SNV or indel R^2 for each SV appears in Figure 1, separated into the following SV types: insertion (INS), duplication (DUP), deletion (DEL), complex structural variant (CPX), and inversion (INV). The median R^2 of the SNV in highest LD with each SV is 0.78 for insertions, 0.13 for duplications, 0.80 for deletions, 0.93 for complex events, and 0.83 for inversions. Similar results hold when samples are subset into genetic ancestry groups (Figure 2). Figure 3 shows the results of stratifying events of each SV type by the genomic sequence context it appears in. The median R^2 value of the SNV in highest LD with each SV, broken into SV types within each genomic sequence context, is given in Table 1. There were no inversions annotated as belonging to simple repeats in the callset. Stratifying by sequence context shows that the low overall LD of duplications was driven by events within SR or SD sequence contexts (median maximum R^2 in SD: 0.15; SR: 0.05), while duplication variants within US or RM contexts have detectable LD comparable to the other SV types (Figure 3; 0.83 median in US, 0.80 median in RM). It should be noted that biological factors, potentially including increased mutation rates and recombination rates in repetitive sequence contexts such as simple repeats and segmental duplications, as well as technical factors such as the difficulty of discovering SVs and SNVs in those contexts, contribute to the expected lower LD scores identified in repetitive regions of the genome. As illustrated in Figure 4, in unique sequence contexts all variant classes have high median LD with a nearby SNV (INS: 0.93; DUP: 0.83; DEL: 0.93; CPX: 0.92; INV: 0.78).

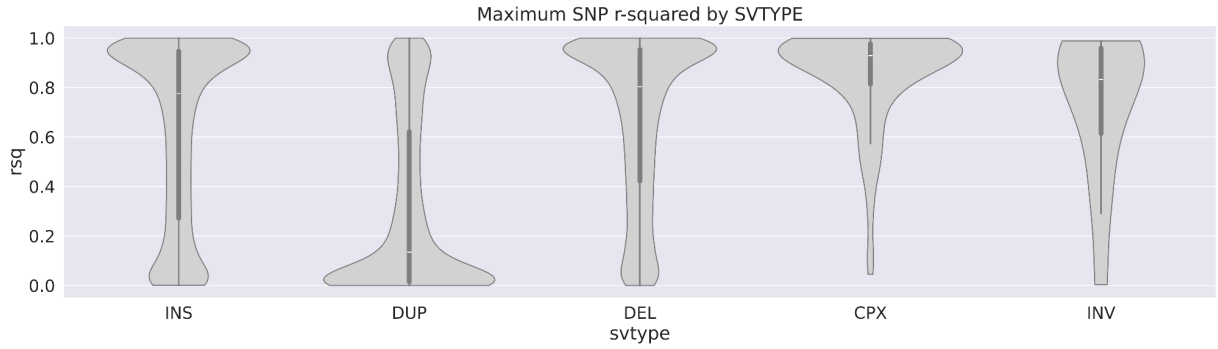

Figure 1 -- The distribution of maximum SNV-SV R^2 values for each SV type in this analysis.

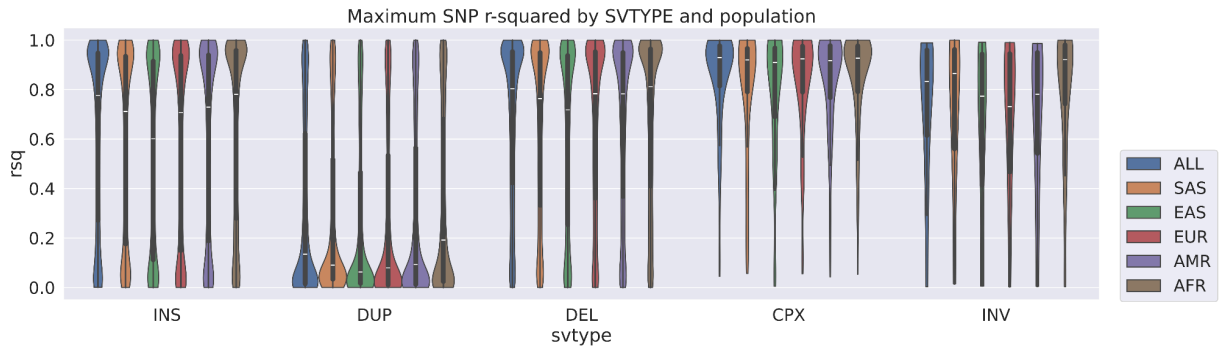

Figure 2 -- The distribution of maximum SNV-SV R^2 values for each SV type, stratified by the genetic ancestry group of the participant (ALL: all samples; EAS: 1KGP-HGDP-EAS-like; AMR: 1KGP-HGDP-AMR-like; AFR: 1KGP-HGDP-AFR-like; SAS: 1KGP-HGDP-SAS-like; EUR: 1KGP-HGDP-EUR-like).

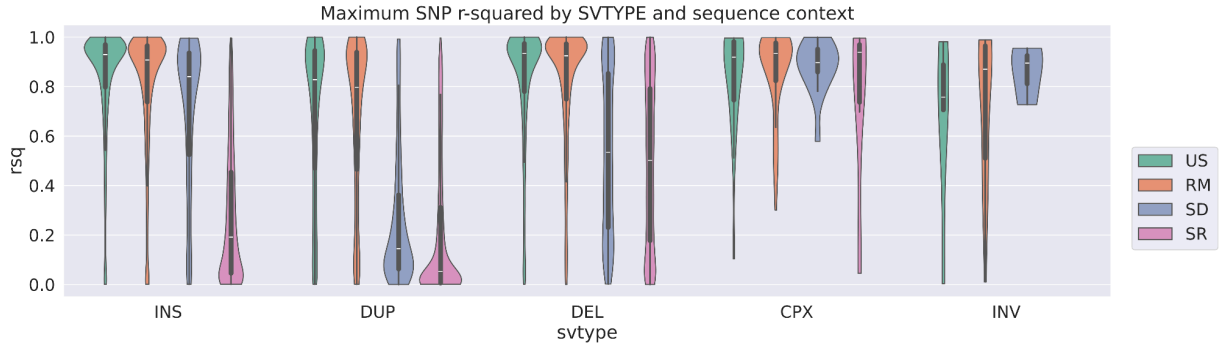

Figure 3 -- The distribution of maximum SNV-SV R^2 values for each SV type, stratified by genomic context (SR: simple repeat; SD: segmental duplication; US: unique sequence; RM: repeat-masked sequence).

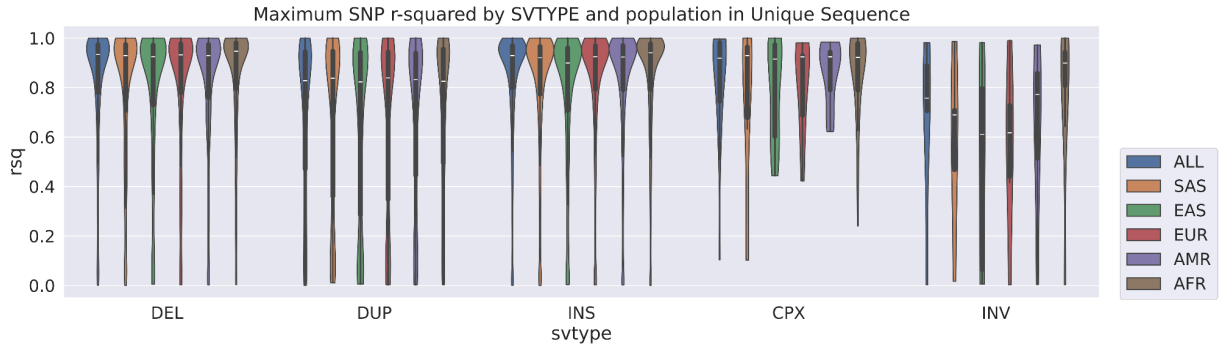

Figure 4 -- The distribution of maximum SNV-SV R^2 values for each SV type when limited to regions of the genome with unique sequence context, stratified by the genetic ancestry group of the participant (ALL: all samples; EAS: 1KGP-HGDP-EAS-like; AMR: 1KGP-HGDP-AMR-like; AFR: 1KGP-HGDP-AFR-like; SAS: 1KGP-HGDP-SAS-like; EUR: 1KGP-HGDP-EUR-like).

Table 1 -- Median SNV-SV R^2 value for each SV type, stratified by genetic ancestry groups and genomic context

		SV Type				
Genetic ancestry group	Sequence Context	DEL	DUP	INS	CPX	INV
ALL	US	0.934	0.828	0.930	0.919	0.757
	RM	0.925	0.797	0.908	0.933	0.870
	SD	0.534	0.145	0.844	0.897	0.895
	SR	0.502	0.052	0.191	0.939	N/A
AFR	US	0.947	0.826	0.941	0.922	0.899
	RM	0.937	0.840	0.920	0.928	0.940
	SD	0.548	0.162	0.844	0.879	0.755
	SR	0.468	0.082	0.202	0.952	N/A
AMR	US	0.929	0.832	0.924	0.924	0.772
	RM	0.915	0.796	0.901	0.921	0.752
	SD	0.483	0.146	0.818	0.918	0.918
	SR	0.477	0.036	0.159	0.839	N/A
EAS	US	0.925	0.823	0.899	0.915	0.611
	RM	0.895	0.765	0.875	0.899	0.773
	SD	0.429	0.147	0.806	0.891	0.886
	SR	0.384	0.029	0.133	0.954	N/A
EUR	US	0.936	0.839	0.925	0.924	0.617
	RM	0.914	0.914	0.905	0.915	0.847
	SD	0.516	0.162	0.830	0.905	0.935
	SR	0.476	0.030	0.149	0.931	N/A
SAS	US	0.930	0.838	0.921	0.929	0.690
	RM	0.911	0.797	0.901	0.921	0.921
	SD	0.474	0.157	0.825	0.905	0.928
	SR	0.445	0.050	0.167	0.843	N/A

Patterns of evolutionary constraint

Methods

Patterns of evolutionary constraint across genes have been previously examined in SNVs and indels and quantified by the loss-of-function observed/expected upper bound fraction (LOEUF) score [8]. Analyses in gnomAD-SV V2 showed that SVs exhibit similar trends of gene-level intolerance to variation [2]. To demonstrate that the CDRv8 srWGS SV callset exhibits the same fundamental biological signals, we replicated the methods in Collins et al. 2020 [2] to examine trends of SV constraint in comparison to SNV constraint. Briefly, we estimated the depletion of rare SVs per gene compared to the expected count of SVs per gene, using a negative binomial regression model.

We subsetted the VCF to sites with a PASS filter status, then to the maximal set of 93,360 unrelated samples in the CDRv8 srWGS SV callset. Next, we computed the number of rare (AF <1%) SVs observed per gene for all autosomal protein-coding genes, across four different classes of functional consequences. The functional consequence categories used in this analysis were predicted loss-of-function (pLOF), copy gain duplication (CG, in which an entire gene is duplicated), intragenic exonic duplication (IED, in which intact exons are duplicated without disrupting coding sequence), and spanning inversion (INV, in which an inversion spans an entire gene).

Based on gene characteristics and these observed counts, we trained a negative binomial regression model to predict the expected counts of SVs of different functional classes for each gene. We incorporated the following factors into the model: gene length, total and median exon length, total and median intron length, number of exons, number of introns, and the proportion of the gene overlapped by segmental duplication regions. We trained the model on the genes exhibiting relatively neutral selection in the 5th to 9th LOEUF deciles. We then applied the model to estimate the expected number of gene-disrupting SVs in each functional category across all autosomal protein-coding genes in GENCODE v39 [9].

We binned genes by LOEUF percentile (resulting in 100 bins containing an average of 189 genes each) and compared the estimated expected counts of rare SVs of each functional class for the genes in each bin to the observed counts. Finally, we used a two-sided Spearman's rank correlation test to assess the correspondence between SV and SNV constraint across all 100 bins of genes.

Results

Figure 5 shows the results of the constraint analysis for rare coding SVs across four different classes of SV functional consequences representing a spectrum of expected impact on the protein. As expected, the depletion of rare pLOF SVs shows the strongest concordance with the depletion of pLOF SNVs as measured by LOEUF (pLOF Spearman correlation test, ρ=0.95, $P<10^{-100}$). There is also a strong relationship between CG SV constraint and LOEUF (CG

Spearman correlation test, ρ=0.90, P<10^{-100}) and a weaker but significant relationship between IED SV constraint and LOEUF (IED Spearman correlation test, ρ=0.73, P<10^{-100}). There is not a significant correlation between INV constraint and LOEUF (INV Spearman correlation test, ρ=0.05, P=6.56x10^{-1}). These results recapitulate the findings in Collins et al. 2020 [2] and show that our findings reflect previously established patterns of evolutionary constraint.

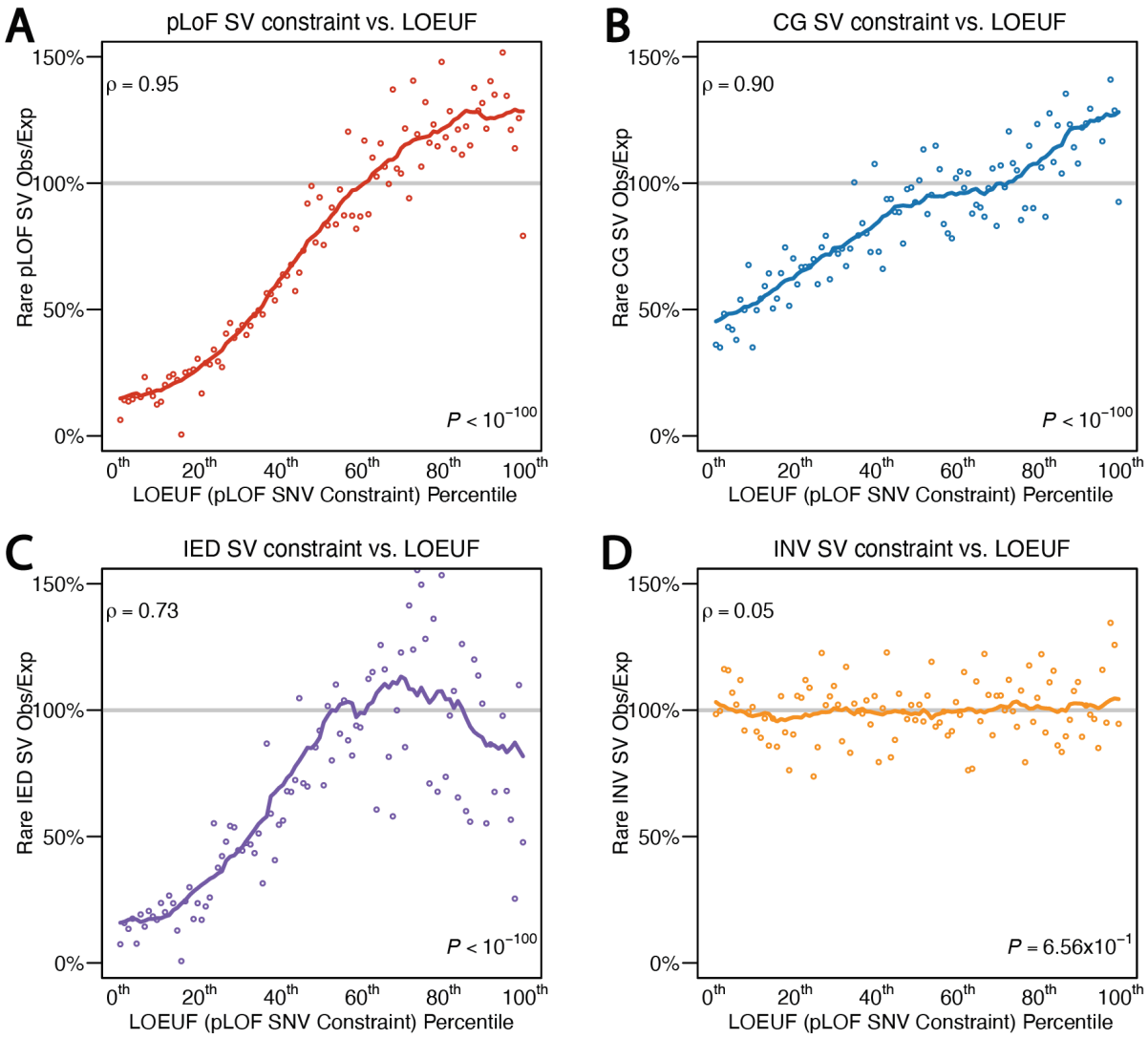

Figure 5 -- Comparing pLOF SNV constraint as measured by LOEUF percentile (x-axis) to binned SV constraint as measured by the rare observed/expected ratio (y-axis) across four different functional classes of SVs: A) predicted loss-of-function SVs (pLOF), B) copy gain duplications SVs (CG), C) intragenic exonic duplication SVs (IED), and D) inversion SVs that span an entire gene (INV). Points represent observed vs. expected SV count ratios for SVs impacting genes in a given LOEUF percentile with a given functional consequence. Solid lines represent 21-point rolling means. The results of the two-sided Spearman correlation test (the correlation ρ and the P-value) are superimposed on each panel.

Comparisons to orthogonal data types

Benchmarking against long-read PacBio sequencing

Data and methods

We evaluated passing non-reference SV genotypes based on evidence derived from long-read whole genome sequencing (lrWGS). The lrWGS SV calls using existing algorithms can confirm SV events with accurate breakpoint resolution, but often miss large insertions and inversions near the lrWGS read size, as well as large copy number variants (CNV) that require read depth evidence to detect. Read depth signatures are used extensively in the GATK-SV short-read pipeline but not in existing lrWGS algorithms. Because of this reduced sensitivity of lrWGS SV calling to large SVs, variants larger than 5 kilobases (kb) were excluded from this analysis.

We performed this analysis on a subset of 97 samples with matched lrWGS data that were held out from training of the GQ filtering model used for refinement of the SV callset (see the Genomic Research Data Quality Report, srWGS SV Genotype Filtering [1]). For each sample, passing non-reference genotypes for eligible variants (SV type DEL, DUP, INS, or INV, with PASS filter status, below 5 kb in length) were assessed against lrWGS using the lrWGS validation tool VaPoR [10] and their overlap with SV calls from lrWGS data from the tools PAV [11], PBSV [12], and sniffles2 [13]. Duplications present a challenge to overlap-based methods of variant matching, as they can be called either as INS or DUP types, with INS calls either at the 5' or 3' end of the duplicated sequence. In order to avoid such complications with variant representation, the evaluated calls were grouped into three main classes prior to variant matching: insertions (encompassing insertions and duplications), deletions, and inversions. SrWGS variants were matched with lrWGS variants of the same comparison class by requiring 10% reciprocal overlap and 50% size similarity. This analysis was performed using the GATK SVConcordance tool [14].

Results

The validation callset generated by GATK-SV included 704,155 total non-reference calls comprising 52,046 unique deletion, duplication, insertion, and inversion variants under 5kb. These calls were strongly supported by lrWGS, with 622,567 (88%) of the PASS genotypes confirmed by at least one lrWGS tool. Figure 6 shows the distributions of support from lrWGS for insertion and deletion SVs, and Figure 7 shows them for inversions. For each intersection, the number of calls is shown with variant size and GQ distributions. Note that the GQ recalibration model was trained on a set of independent samples using lrWGS support criteria. Therefore, a higher GQ reflects that the call was similar to calls in the training set with support from VaPoR and at least one of the three lrWGS SV algorithms (see the Genomic Research Data Quality Report, srWGS SV Genotype Filtering [1]).

There was a high degree of consensus among the lrWGS callers, with only 58,705 (9.4% of confirmed) srWGS SV calls supported by just one lrWGS SV caller and 525,380 (84%) supported by at least three. Calls with no lrWGS support had overall lower genotype quality (GQ) scores (median 49) compared to supported calls (median 89), which is consistent with expectations. Notably, PBSV was the most consistent with srWGS SV calls from GATK-SV, supporting 583,802 (94% of confirmed) srWGS calls with a median GQ of 89, compared to the remaining 38,765 lrWGS supported calls with a median GQ of 58.

The distribution of calls produced by the three non-depth based srWGS SV calling tools used by GATK-SV (Manta [15], Wham [16], and MELT [17]) and the fraction of calls with lrWGS support for each is shown in Figure 6B. Overall, Manta produced 569,235 (81%) of passing calls, 89% of which were supported by at least one lrWGS SV discovery method. In addition, MELT contributed 139,387 (20%) of the calls with 88% lrWGS support. Wham is utilized in this pipeline to access a subset of small duplications that are missed by other algorithms, with 93,355 variants retained after filtering (13% of total) and 87% lrWGS support. Similar to insertion and deletion SVs, inversions exhibited a high degree of support from lrWGS, with 805 of 833 (97%) validated (Figure 7).

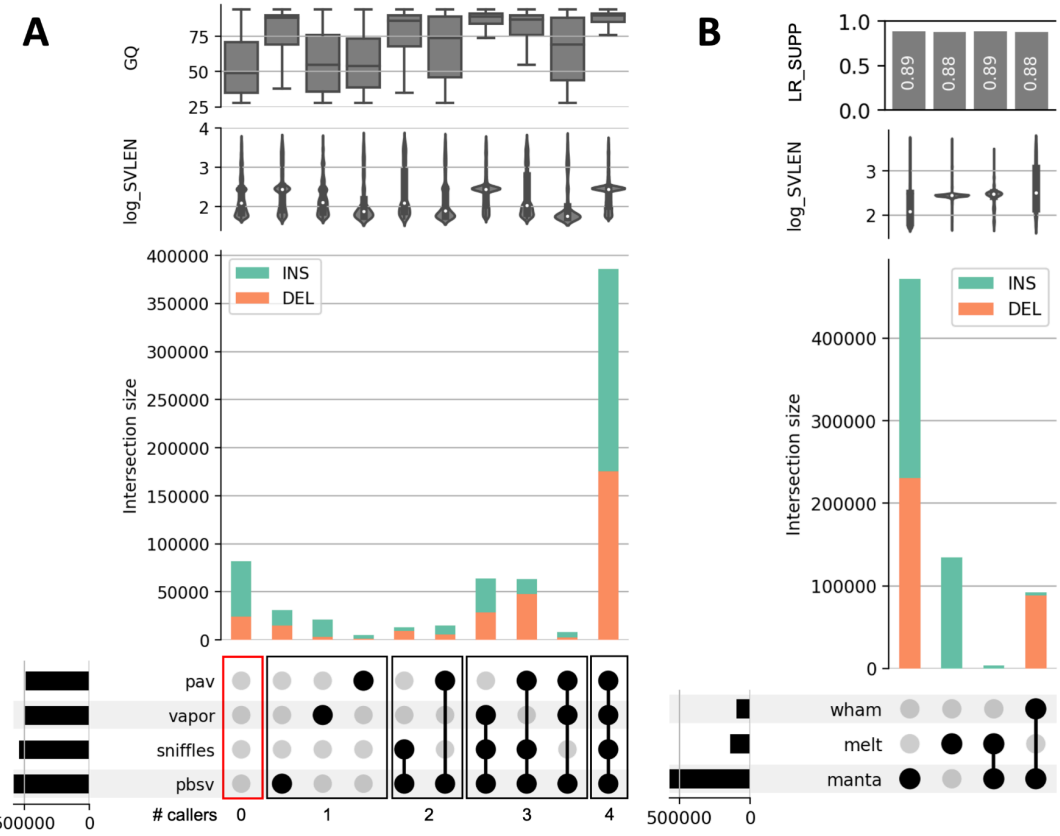

Figure 6 -- Evaluation of passing srWGS insertion (INS) and deletion (DEL) calls under 5 kb against lrWGS tools. The insertion class encompasses all sequence gain events, including duplications. (A) Insertion and deletion SV calls supported by each lrWGS tool (PAV, VaPoR,

sniffles2, and PBSV). Filled circles indicate combinations of tools that support the call counts in each column. Combinations with fewer than 5,000 total calls are omitted for clarity. Violin plots of genotype quality and $\log_{10}$ of variant length distributions are superposed over each combination. Total supported calls for each lrWGS tool are plotted at the bottom-left. (B) Insertion and deletion SV calls supported by each srWGS SV caller. The top panel shows the fraction of calls with support from at least one lrWGS tool. Combinations with fewer than 2,000 calls are omitted.

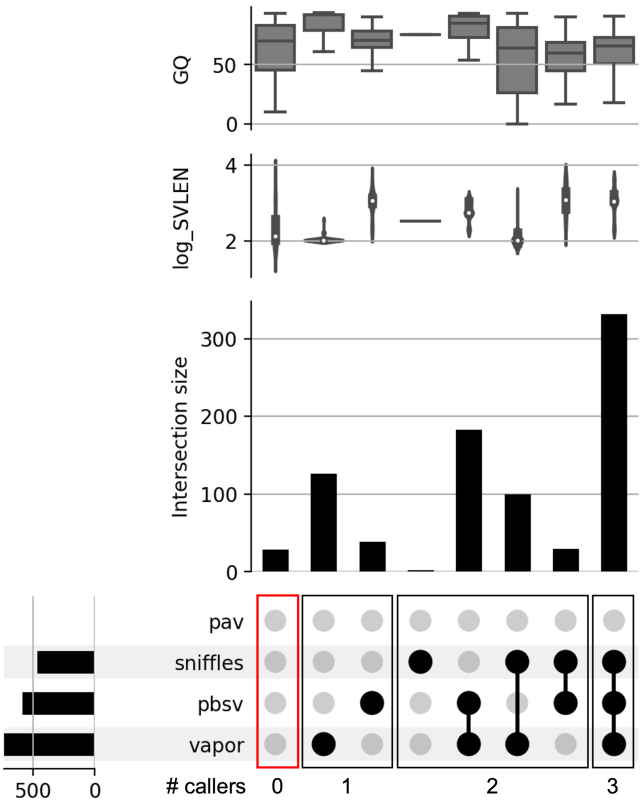

Figure 7 -- Evaluation of passing srWGS inversion calls under 5 kb against lrWGS tools (PAV, sniffles2, PBSV, and VaPoR). Filled circles indicate combinations of tools that support the call counts in each column. All non-empty combinations are shown. Violin plots of genotype quality and $\log_{10}$ of variant length distributions are superposed over each combination. Total supported calls for each lrWGS tool are plotted at the bottom-left.

Benchmarking large CNVs against microarrays

Data and Methods

In a randomly-selected subset of 10,000 samples, we evaluated all deletions and duplications greater than 10 kb and less than 10 Mb in length on the autosomes using array intensity data

from the LRR field of the array VCFs (available on the Researcher Workbench and described in 'How the *All of Us* Genomic Data are Organized'). To conduct this evaluation we used the GenomeSTRiP IntensityRankSumAnnotator (IRS) tool [18,19]. The IRS tool compares the array probe intensity values between samples predicted to carry the CNV and those predicted to be non-carriers (according to genotypes in the SV VCF), using all probes that are within the CNV interval. Using a non-parametric test, the IRS tool assigns a p-value to each CNV which indicates if the CNV genotypes are supported by the intensity data. In addition to using site-level p-values, the authors of the test recommend [18,19] using IRS to calculate a callset level false discovery rate (FDR) by computing $2 * \frac{M}{N}$, where *M* is the number of sites where the IRS p-value is greater than 0.5 and *N* is the total number of sites. CNVs greater than 10 Mb were excluded due to the computational requirements required to evaluate array concordance using these methods and the fact that the majority of these large events are likely to be somatic events that can be challenging to confirm. We note that performance of this validation can be compromised for smaller CNVs if there is insufficient probe density in the CNV region on the microarrays, which is a common challenge for CNVs less than 20 kb in size. Nonetheless, this validation can still be informative in the 10-20 kb size range for many regions of the genome.

We ran the IRS test on all samples at each site. The IRS test requires that an intensity value be present for all samples. Therefore, if a sample had a missing data value for one or more of the probes covered by the CNV interval, we set the intensity value to a random value such that the rank of the inserted value within the cohort would be uniformly distributed. This was achieved by choosing another sample at random from the set of samples with non-missing values for that probe and setting the missing sample's intensity value to that of the randomly chosen sample. The substitution of missing data points with randomly chosen values was necessary for testing the callset against the entire cohort, but could inflate the FDR estimates provided by the IRS test. We excluded probes for which over half of the samples had missing values; this resulted in the exclusion of one probe from the evaluation.

Results

After removing 1,624 CNV sites which did not overlap any array probes and could not be tested, 34,510 autosomal CNVs of size 10 kb to 10 Mb were evaluated using this test, including 20,274 deletions and 14,236 duplications. Among the 20,274 deletions evaluated, 210 had an IRS p-value greater than 0.5, resulting in an estimated FDR of 2.07% for all deletions tested using the callset-wide evaluation procedure described above. We also analyzed the results with a stricter p-value cutoff of 0.01, which was the threshold used to select sites for molecular validation based on IRS results in a previous study [18]. With the stricter cutoff, 93.4% of deletions validated. The results for deletions in different size ranges are shown in Table 2.

Out of the 14,236 duplications evaluated, 405 had a p-value over 0.5, resulting in an estimated callset FDR of 5.69% for duplications. Using the stricter p-value threshold of 0.01, 89.8% of duplications validated. Duplication results by size range are shown in Table 3. As in the discussion of array benchmarking for the CDRv7 release [4], we note that 7.0% (298 / 4,249) of duplications and 5.9% (541 / 9,108) of deletions between 10 kb and 20 kb span only one probe,

reducing the statistical power of the IRS test to validate these events at the p-value < 0.01 level. Overall, these results show that large CNVs in this callset were strongly supported by microarrays, with a very low estimated FDR for both large deletions and large duplications.

Table 2 -- Deletion SV array validation results, stratified by SV size

Size range	Number of sites included	Estimated Callset FDR	P-value < 0.01
10-20kb	9,108	2.66%	8,152 (89.5%)
20-50kb	5,390	1.37%	5,161 (95.8%)
50-100kb	2,699	1.93%	2,610 (96.7%)
100kb-1Mb	2,897	1.31%	2,851 (98.4%)
1-10Mb	75	0%	75 (100%)

Table 3 -- Duplication SV array validation results, stratified by SV size

Size range	Number of sites included	Estimated Callset FDR	P-value < 0.01
10-20kb	4,249	7.44%	3,456 (81.3%)
20-50kb	4,072	5.21%	3,708 (91.1%)
50-100kb	2,252	5.33%	2,122 (94.2%)
100kb-1Mb	3,413	4.16%	3,288 (93.3%)
1-10Mb	143	1.40%	142 (99.3%)

References

[1] "*All of Us* Genomic Quality Report" *All of Us Research Program*, https://support.researchallofus.org/hc/en-us/articles/29390274413716-All-of-Us-Genomic-Quality-Report

[2] Collins, R.L., Brand, H., Karczewski, K.J. *et al.* A structural variation reference for medical and population genetics. *Nature* **581**, 444-451 (2020). https://doi.org/10.1038/s41586-020-2287-8

[3] Byrska-Bishop, Marta et al. "High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios." *Cell* vol. 185,18 (2022): 3426-3440.e19. doi:10.1016/j.cell.2022.08.004

[4] "Benchmarking and quality analyses on the All of Us v7 short read structural variant calls", *All of Us Research Program*,

https://support.researchallofus.org/hc/en-us/articles/14941865780500-Benchmarking-and-quality-analyses-on-the-All-of-Us-short-read-structural-variant-calls-ARCHIVED
[5] Hill, W G, and A Robertson. “Linkage disequilibrium in finite populations.” *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik* vol. 38,6 (1968): 226-31. doi:10.1007/BF01245622
[6] Zhao, Xuefang et al. “Expectations and blind spots for structural variation detection from long-read assemblies and short-read genome sequencing technologies.” *American journal of human genetics* vol. 108,5 (2021): 919-928. doi:10.1016/j.ajhg.2021.03.014
[7] ***Structural Variants*** (n.d.). Retrieved March 3, 2023, from https://gatk.broadinstitute.org/hc/en-us/articles/9022476791323-Structural-Variants
[8] 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
[9] Frankish A, Diekhans M, *et al.* GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 2019 Jan 8;47(D1):D766-D773. doi: 10.1093/nar/gky955. PMID: 30357393; PMCID: PMC6323946.
[10] Zhao X, Weber AM, Mills RE. A recurrence-based approach for validating structural variation using long-read sequencing technology. Gigascience. 2017 Aug 1;6(8):1-9. doi: 10.1093/gigascience/gix061. PMID: 28873962; PMCID: PMC5737365.
[11] P. Ebert, P. A. Audano, Q. Zhu et al., Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021).
[12] **PacBio structural variant calling and analysis tools (PBSV)**, Retrieved March 3, 2023, from https://github.com/PacificBiosciences/pbsv.
[13] Sedlazeck FJ, Rescheneder P, Smolka M, et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods*. 2018 Jun;15(6):461-468. doi: 10.1038/s41592-018-0001-7. Epub 2018 Apr 30. PMID: 29713083; PMCID: PMC5990442.
[14] GATK Team “SVConcordance (Beta) – GATK.” *GATK*, 20 Mar. 2023, https://gatk.broadinstitute.org/hc/en-us/articles/13832773767963-SVConcordance-BETA-.
[15] Chen, X. *et al.* (2016) Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*, 32, 1220-1222. doi:10.1093/bioinformatics/btv710
[16] Kronenberg ZN, Osborne EJ, Cone KR, Kennedy BJ, Domyan ET, Shapiro MD, *et al.* (2015) Wham: Identifying Structural Variants of Biological Consequence. PLoS Comput Biol 11(12): e1004572. https://doi.org/10.1371/journal.pcbi.1004572
[17] Gardner, E. J., Lam, V. K., Harris, D. N., Chuang, N. T., Scott, E. C., Mills, R. E., Pittard, W. S., 1000 Genomes Project Consortium & Devine, S. E. The Mobile Element Locator Tool (MELT): Population-scale mobile element discovery and biology. *Genome Research*, 2017. **27**(11): p. 1916-1929.
[18] Mills, Ryan E et al. Mapping copy number variation by population-scale genome sequencing. Nature vol. 470,7332 (2011): 59-65. doi:10.1038/nature09708
[19] Handsaker, R., Van Doren, V., Berman, J. *et al.* Large multiallelic copy number variations in humans. *Nat Genet* **47**, 296-303 (2015). https://doi.org/10.1038/ng.3200