Benchmarking and quality analyses on the *All of Us* v7 short read structural variant calls

Introduction

Widespread benchmarking of structural variants (SVs) from short read whole genome sequencing (srWGS) remains challenging, often due to lack of orthogonal data for comparison. The *All of Us* cohort of samples with srWGS SV data is somewhat unique in the availability of matched genomic datasets (i.e SNP arrays, srWGS SNPs and Indels, and long read genome sequencing [lrWGS]). There also exists a number of intrinsic measures that can be used to assess the technical quality of a dataset (e.g Hardy-Weinberg Equilibrium). Combining these methods, we have a unique opportunity for a high-quality assessment of SV generated via srWGS. Overall, we assess 7 measures of technical quality for the GATK-SV *All of Us* dataset, described in the main QC report and this supplemental document.

In the Structural Variant QC Results section of 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 benchmarking report, we additionally describe:

4. Linkage disequilibrium with srWGS SNPs and Indels
5. Patterns of evolutionary constraint
6. Benchmarking against long read sequencing data
7. Benchmarking against microarrays

In addition to these QC analyses, in this report we describe an analysis to benchmark the performance of the DRAGEN 3.4.12 aligner compared to BWA for the discovery of SVs with GATK-SV.

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 SNVs and Indels, the presence of nearby SNVs in linkage disequilibrium (LD) with our 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 v0.2.107 in a Python notebook backed by a Spark 2.4.5 cluster. LD analyses were conducted for the full cohort as well as subsets of the cohort that shared an assigned super-population ancestry and contained at least 1,000 samples. The ancestry categories chosen under these criteria were European (EUR; n=4,691), African (AFR; n=4,176), and admixed American (AMR; n=1,430). We analyzed LD between all SVs with PASS filter status and SNPs/indels with PASS filter status that had a minor allele frequency of at least 1% in either the full cohort or one of these population subsets.

LD between the callsets was computed 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 population subset, 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 SNPs and indels after minor allele frequency filtering and B_{ij} is the number of alternate alleles for sample j at SNP/indel site i.

We defined LD as the R^2 of alternate allele dosage between each pair consisting of one SV site and one SNP site [4]. 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/SNP pairs where the SNP 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 SNPs located within the window using Hail's block matrix sparsification functionality. For each SV we identified the SNP 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 [5].

Results

A violin plot of the maximum SNP or indel R^2 for each SV appears in Figure 1, broken out by SV type. The median R^2 of the SNP in highest LD with each SV is over 0.7 for all SV types, except duplications. Similar results hold when samples are subset into sub-populations (Figure 2). There were no inversions annotated as belonging to simple repeats in the callset. The median R^2 value of the SNP in highest LD with each SV, broken into SV types and each genomic sequence context, is given in Table 1. Stratifying by sequence context shows that duplications within SR or SD sequence contexts have lower SNP LD than those in US or RM contexts (Figure 3). 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 SNPs in those contexts, contribute to the expected lower LD scores identified in repetitive regions of the genome.

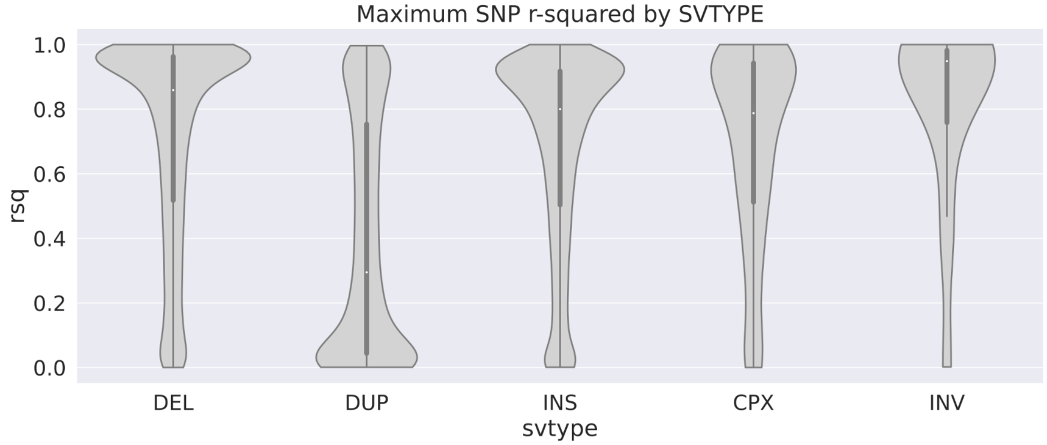

Figure 1 – The distribution of maximum SNP-SV R^2 values for each SV type. The SV types in this analysis were: deletion (DEL), duplication (DUP), insertion (INS), complex event (CPX), and inversion (INV).

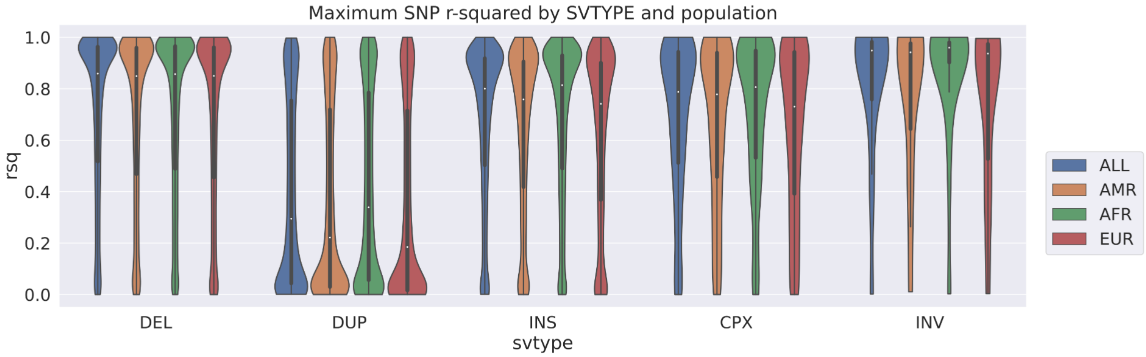

Figure 2 – The distribution of maximum SNP-SV R^2 values for each SV type, stratified by predicted sample ancestry (ALL: all samples; EUR: European; AMR: Admixed American; AFR: African).

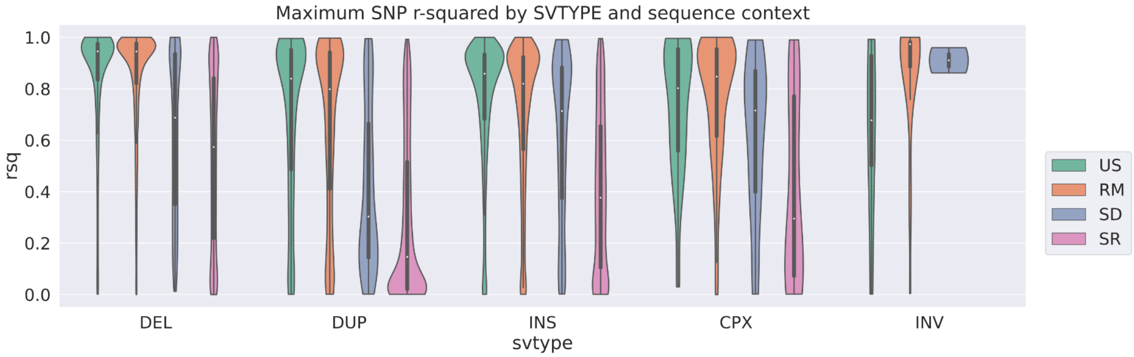

Figure 3 – The distribution of maximum SNP-SV R^2 values for each SV type, stratified by genomic context. (SR: Simple Repeat; SD: Segmental Duplication; US: Unique Sequence; RM: Repeatmasked sequence)

Table 1 – Median SNP-SV R^2 value for each SV type, stratified by ancestry and genomic context

		SV type				
Population	**Sequence Context**	**DEL**	**DUP**	**INS**	**CPX**	**INV**
ALL	**US**	0.945	0.840	0.859	0.802	0.677
	RM	0.945	0.799	0.820	0.847	0.974
	SD	0.687	0.303	0.714	0.715	0.911
	SR	0.573	0.146	0.376	0.295	N/A
AFR	**US**	0.951	0.845	0.870	0.868	0.925

	RM	0.947	0.825	0.834	0.856	0.967
	SD	0.709	0.368	0.748	0.644	0.824
	SR	0.534	0.175	0.349	0.481	N/A
EUR	**US**	0.948	0.852	0.829	0.727	0.467
	RM	0.940	0.737	0.773	0.836	0.943
	SD	0.695	0.327	0.645	0.721	0.945
	SR	0.557	0.081	0.297	0.253	N/A
AMR	**US**	0.943	0.825	0.833	0.813	0.500
	RM	0.938	0.765	0.788	0.854	0.958
	SD	0.614	0.310	0.664	0.717	0.915
	SR	0.562	0.101	0.328	0.319	N/A

Patterns of evolutionary constraint

Methods

Patterns of evolutionary constraint across genes have been previously examined in SNVs and indels [6], and analyses in gnomAD V2 showed that SVs exhibit similar trends of gene-level intolerance to variation [2]. To demonstrate that the v7 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 the maximal set of 11,306 unrelated samples in the v7 srWGS SV callset, then 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). Next, we trained the model to predict the expected counts of SVs of different functional classes for each gene based on factors like gene length, number of exons and introns, and overlap with segmental duplication regions. In order to estimate the expected number of SVs per gene under neutral selection, the model was trained on genes in the 5th-9th deciles for the loss-of-function observed/expected upper bound fraction

(LOEUF), a metric for constraint against rare pLOF SNVs [6]. We then applied the model to estimate expected counts of SVs in each functional class across all autosomal protein-coding genes.

We then binned genes by LOEUF percentile (resulting in 100 bins containing an average of 170 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.

Since the LOEUF values for each gene were computed using hg19 data but the SVs were annotated using the MANE Select Plus Clinical GTF (v0.95), we mapped gene symbols from hg19 to hg38 (GENCODE release 33) using Supplementary Table 7 provided in Fu et al. 2022 [7].

Results

Figure 4 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=0.94, $P<10^{-100}$). There is also a strong relationship between CG SV constraint and LOEUF (CG Spearman correlation test, ρ=0.75, $P<10^{-100}$) and a weaker but significant relationship between IED SV constraint and LOEUF (IED Spearman correlation test, ρ=0.64, $P=4.85\times10^{-13}$). There is not a significant correlation between INV constraint and LOEUF (INV Spearman correlation test, ρ=0.12, $P<2.24\times10^{-1}$). These results recapitulate the findings in Collins et al. 2020 [2] and show that our findings reflect previously established patterns of evolutionary constraint.

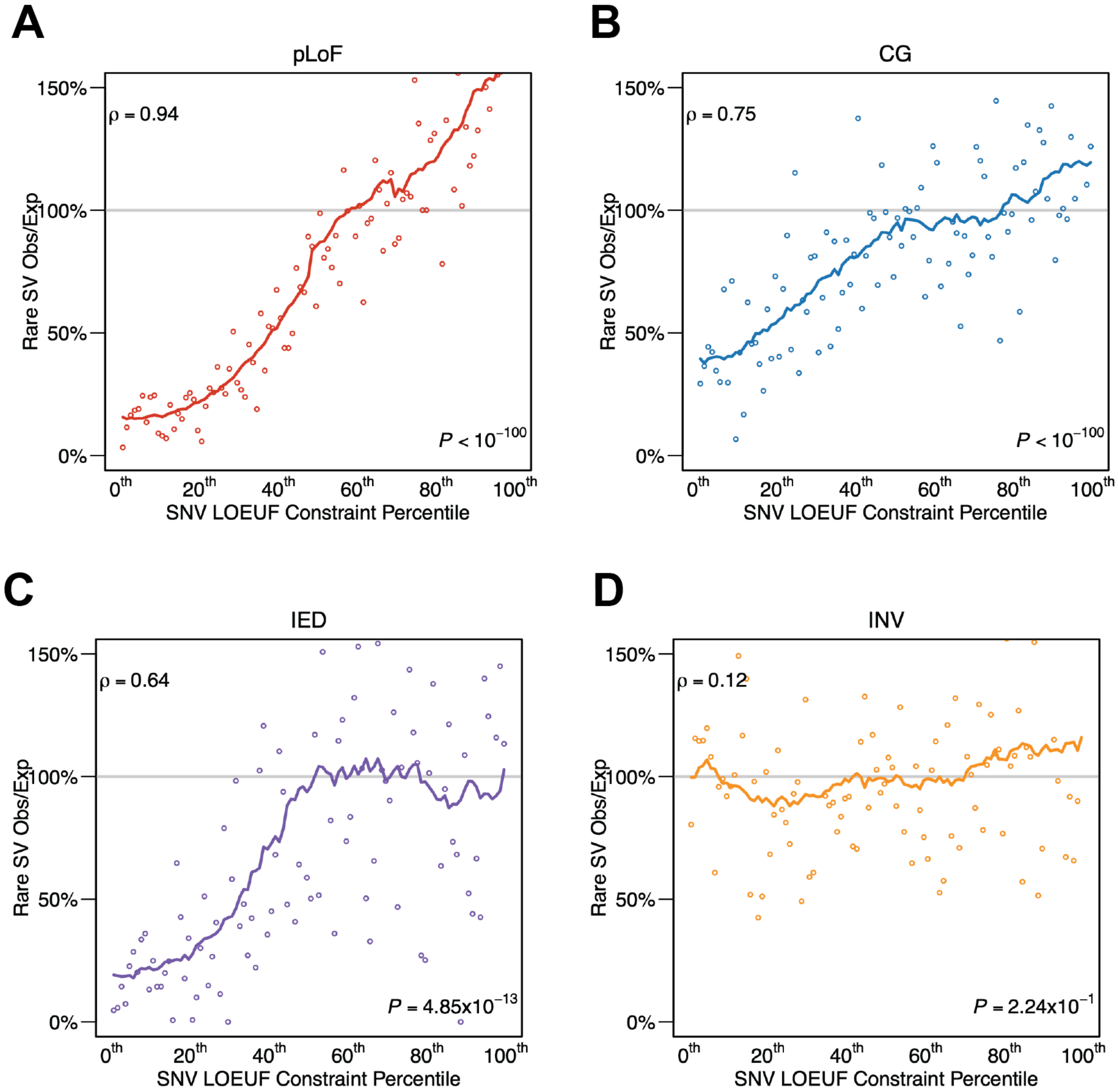

Figure 4 – Comparing pLOF SNV constraint to binned SV constraint in four different SV functional classes: A) predicted loss-of-function (pLOF), B) copy gain duplications (CG), C) intragenic exonic duplications (IED), and D) inversions that span an entire gene (INV). Points represent binned observed vs. expected SV count ratios compared to the LOEUF percentile from SNVs. 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 lrWGS. Long read SV calls using existing algorithms are ideal for confirmation of SV events with accurate breakpoint resolution, but are not sensitive to large insertions and inversions near the lrWGS read size nor to large copy number variants (CNV) that must be detected by read depth signatures. 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 67 samples with matched lrWGS data that were held out from training of the GQ filtering model used for refinement of the SV callset (see srWGS SV Genotype Filter section of the Genomic Research Data Quality Report [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 [8] and their overlap with SV calls from lrWGS data from the tools PAV [9], PBSV [10], and sniffles [11]. 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: gains (DUP and INS), losses (DEL), and inversions prior to variant matching. 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 [12].

Results

The validation callset generated by GATK-SV included 494,147 total non-reference calls comprising 40,668 unique DEL, DUP, INS, and INV variants. These calls were strongly supported by lrWGS, with 445,859 (90%) of the PASS genotypes confirmed by at least one lrWGS tool. Figure 5 shows the distributions of support from lrWGS for gain and loss SVs, and Figure 6 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 srWGS SV Genotype Filter section of the Genomic Research Data Quality Report [1]).

There was a high degree of consensus among the lrWGS callers, with only 35,672 (8.0% of confirmed) srWGS SV calls supported by just one lrWGS SV caller and 372,636 (84%)

supported by at least three. Calls with no lrWGS support had overall lower genotype quality (GQ) scores (median 43) 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 428,333 (96% of confirmed) srWGS calls with a median GQ of 89, compared to the remaining 17,526 lrWGS supported calls with a median GQ of 57.

The distribution of calls produced by the three non-depth based srWGS SV calling tools used by GATK-SV (Manta [13], Wham [14], and MELT [15]) and the fraction of calls with lrWGS support for each is shown in Figure 5B. Overall, Manta produced 444,139 (90%) of passing calls, 93% of which were supported by at least one lrWGS SV discovery method. In addition, MELT contributed 114,827 (23%) of the calls with 85% lrWGS support. Note that while only 69% of calls unique to MELT validated with lrWGS in the final call set, applying a more stringent GQ filter cutoff of 35 for mobile element INS events results in a 91% validation rate overall for MELT while losing less than 0.6% sensitivity to calls under 5 kb. While Wham made only 261 unique calls, it contributed 42,391 (8.6%) in total with 86% lrWGS support. Similar to gain and loss SVs, inversions exhibited a high degree of support from lrWGS, with 524 of 542 (97%) supported by at least one tool and 377 (70%) supported by three (Figure 6).

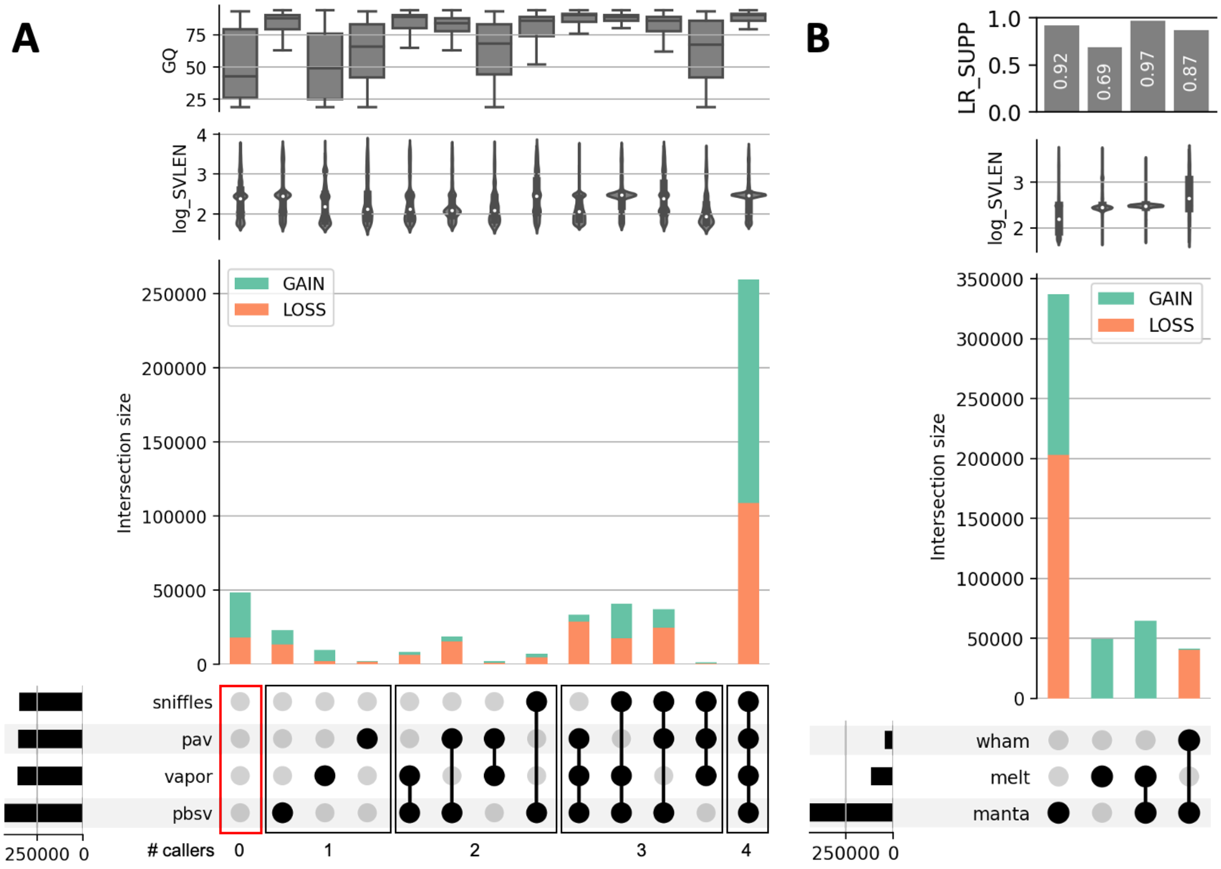

Figure 5 – Evaluation of passing srWGS gain and loss calls under 5 kb against lrWGS tools. (A) Distribution of lrWGS tool support for gain and loss SV classes. Filled circles indicate combinations of tools that support the call counts in each column (combinations with fewer than

1,000 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) Distribution of srWGS tool support. Top panel shows the fraction of calls with support from at least 1 lrWGS tool.

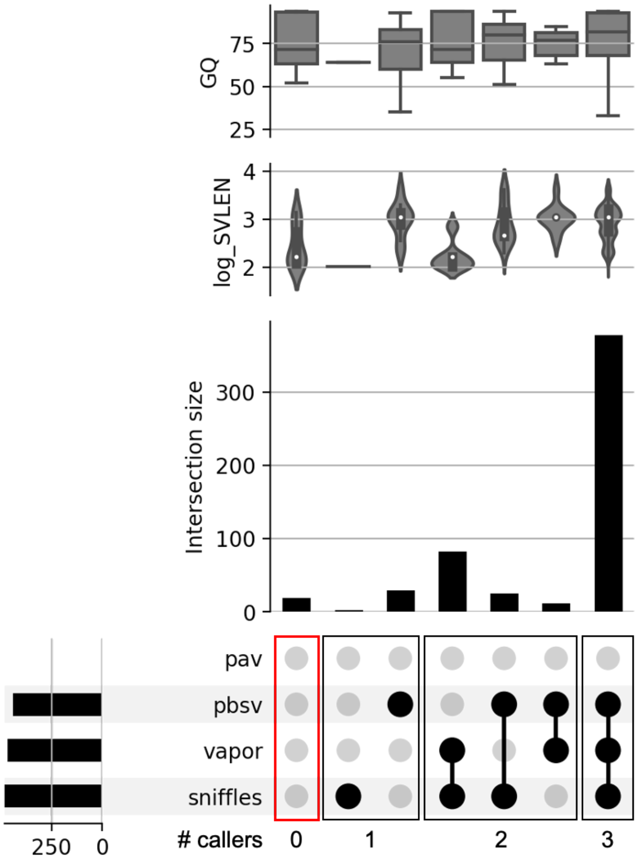

Figure 6 – Evaluation of passing srWGS inversion calls under 5 kb against lrWGS tools. Data is plotted as in Figure 5 for gain and loss calls but with all non-empty combinations shown.

Benchmarking large CNVs against microarrays

Data and Methods

We evaluated all deletions and duplications greater than 10 kb 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 [16,17]. 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 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.

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.

Results

After removing 1,587 duplication sites which did not overlap any array probes and could not be tested, 29,133 autosomal CNVs of size 10kb or greater were evaluated using this test, including 16,867 deletions and 12,265 duplications. 67 out of 16,867 deletions had an IRS p-value greater than 0.5, resulting in an estimated FDR of 0.79% for all deletions tested using the callset-wide evaluation procedure described above. 96.7% of deletions were validated using a more stringent 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 [16]. The results for deletions in different size ranges are shown in Table 2.

Table 2 – Array validation results for deletions in different size ranges

	10kb-20kb	**20-50kb**	**50-100kb**	**100kb-1Mb**	**>1Mb**
Sites	7753	4620	2136	2188	62
Estimated Callset FDR	0.85%	0.74%	0.66%	0.37%	0%
P-value < 0.01	7328 (94.5%)	4545 (98.4%)	2121 (99.3%)	2173 (99.3%)	62 (100%)

Out of the duplications evaluated, 89 had a p-value over 0.5, resulting in an estimated callset FDR of 1.45%. 94.1% of duplications validated at the 0.01 p-value threshold. Duplication results by size range are shown in Table 3. We note that 6% (221 / 3591) of duplications (221 / 3591) and deletions (411 / 7342) between 10kb and 20kb 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 3 – Array validation results for duplications in different size ranges

	10kb-20kb	**20-50kb**	**50-100kb**	**100kb-1Mb**	**>1Mb**
Sites	3591	3525	2098	2820	138
Estimated Callset FDR	2.28%	1.36%	1.14%	0.57%	0%
P-value < 0.01	3094 (86.2%)	3501 (99.3%)	2067 (98.5%)	2801 (99.3%)	138 (100%)

Comparing the BWA and DRAGEN aligners for structural variant calling

Background

The Burrows-Wheeler Aligner (BWA) [18] has remained the field standard for sequence alignment over the past decade and has been the tool of choice for most large-scale sequencing studies to date (e.g. gnomAD, TopMED) [19,20]. Recently, Illumina developed the DRAGEN Aligner which has shown a slight but noticeable improvement for short variant (SNV, indel) calling when compared to BWA [21]. An equivalent comparison looking at DRAGEN vs. BWA for structural variants has yet to be performed and is critical to ensure we can accurately detect SVs using the DRAGEN aligner. In the following analysis we compare 161 samples from the 1000 Genomes Project (1KGP) [3] that have been aligned with both BWA-MEM 0.7.15 and DRAGEN 3.4.12. We included the 23 1KGP samples with matched long read Pacific Biosciences (PacBio) sequences and SV calls [9] to allow for benchmarking against orthogonal data. We then apply GATK-SV on each aligned file and compare the results across aligners.

Data and Methods

Experimental setup

The 161 1KGP samples were aligned with BWA-MEM 0.7.15 as described in the recent 1KGP study [3]. We realigned these sequences with DRAGEN 3.4.12 using the *All of Us* DRAGEN 3.4.12 GRCh38 specifications. We then applied GATK-SV on both alignments using identical settings on Terra. See Appendix A for additional technical details. Downstream filtering and refinement steps were not applied because we did not have equivalent methods available for BWA and DRAGEN SVs. In particular, the genotype filtering method used for the *All of Us* v7 srWGS SV callset was trained and applied on DRAGEN-aligned data and may perform differently on BWA-aligned data.

Comparison methods

SVs from data aligned with DRAGEN (DRAGEN SVs) and BWA (BWA SVs) were compared using custom scripts. We considered a pair of SVs from different aligners to be overlapping events if they shared the same SV type and met one of the following criteria:

1. Deletions and duplications under 5 kb sharing a minimum of 10% reciprocal overlap
2. Deletions and duplications over 5 kb sharing a minimum of 50% reciprocal overlap
3. Insertions having breakpoints within 100 base pairs (bp) of each other

In addition, we allowed SVs of different SV types to match under the following circumstances:

4. Insertions and duplications could match if their breakpoints were within 100 bp
5. Inversions and complex SVs could match if the intervals covered by the SV events had at least 50% reciprocal overlap

Results

Comparison of SV sites

Decent overlap was observed between SV sites aligned with BWA and DRAGEN: 88.8% of DRAGEN SVs overlapped BWA SVs and 87.5% of BWA SVs overlapped DRAGEN SVs (Figure 7A). The concordance, i.e. the proportion of SVs that were shared by SVs aligned with both methods, was not uniform across the genome; SVs in highly repetitive SD and SR regions, the genomic regions that are well documented to have limitations for short reads to align and detect SVs [22,23], had lower overlap (75-89%) than RM and US (94-97%, Figure 7B). Additionally, the overlap patterns differed across SV types: duplications and complex SVs, which are more challenging for current short read algorithms than other SV types [24], demonstrated lower concordance between aligners than deletions, insertions, and inversions (Figure 7C-D) .

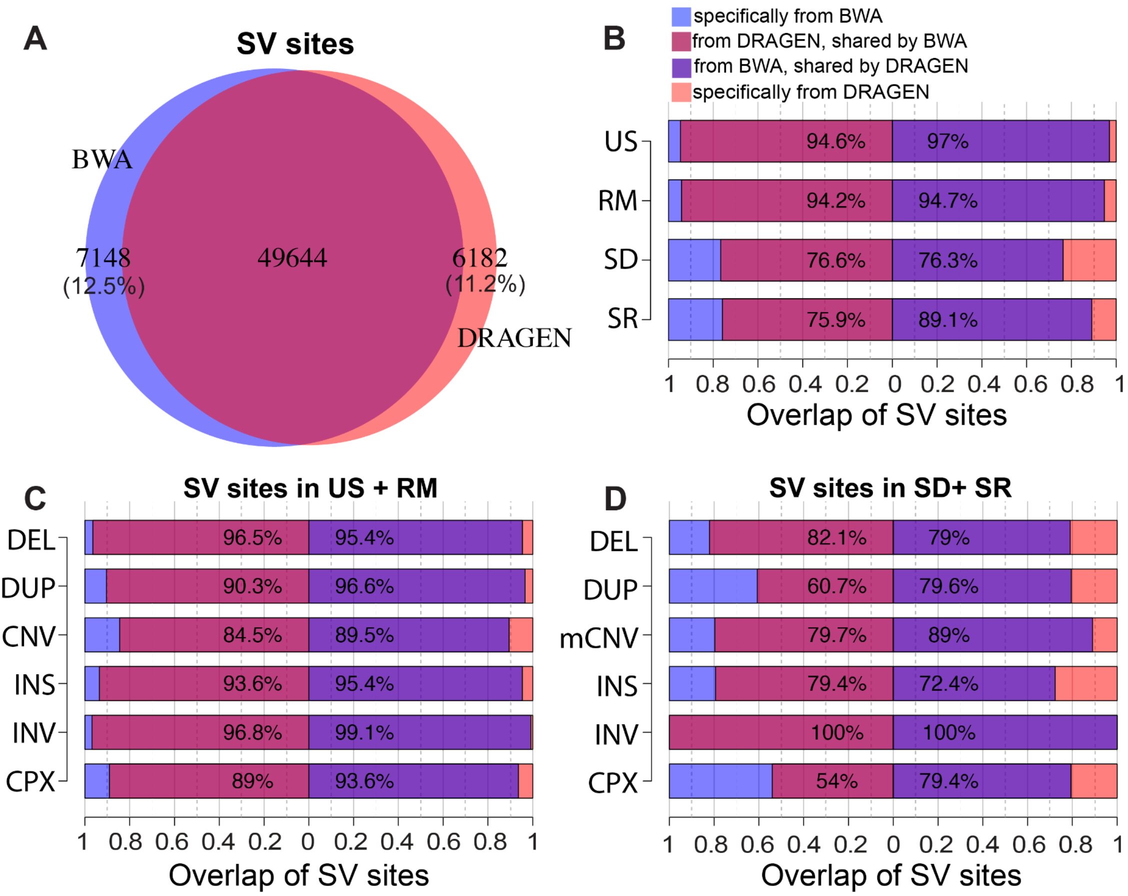

Figure 7 – Overlap between DRAGEN and BWA SVs. A. Overlap of SV sites between BWA and DRAGEN callsets. B. Overlap of SVs in different genomic contexts between BWA and DRAGEN callsets. C-D. Overlap of each SV type between BWA and DRAGEN callsets in C) less repetitive RM and US regions and D) highly repetitive SD and SR sequences.

Comparison of SV genotypes

The 23 samples with matched lrWGS data were used for genotype-level comparisons between the BWA and DRAGEN SVs. When comparing the same sample between the two callsets, 86.8% of DRAGEN SV genotypes match a BWA SV genotype and 83.7% of BWA SV genotypes match a DRAGEN SV genotype. Consistent with site-level observations, SVs in the highly repetitive SD and SR sequences had lower overlap between aligners than RM and US sequences, and duplications had lower overlap than deletions and insertions (Figure 8A). When restricting to SVs that were validated in matched lrWGS data by VaPoR [8] or overlap with lrWGS SV calls [9], higher overlap was observed between the aligners. This increase is consistent across genomic contexts and SV types (Figure 8B). This suggests high-confidence SVs that validate with lrWGS data are more likely to be found from both aligners. As expected, we observed higher concordance between aligners when an SV was discovered by multiple algorithms rather than a single algorithm (Figure 8C), or was supported by multiple types of alignment evidence rather than a single evidence type (Figure 8D).

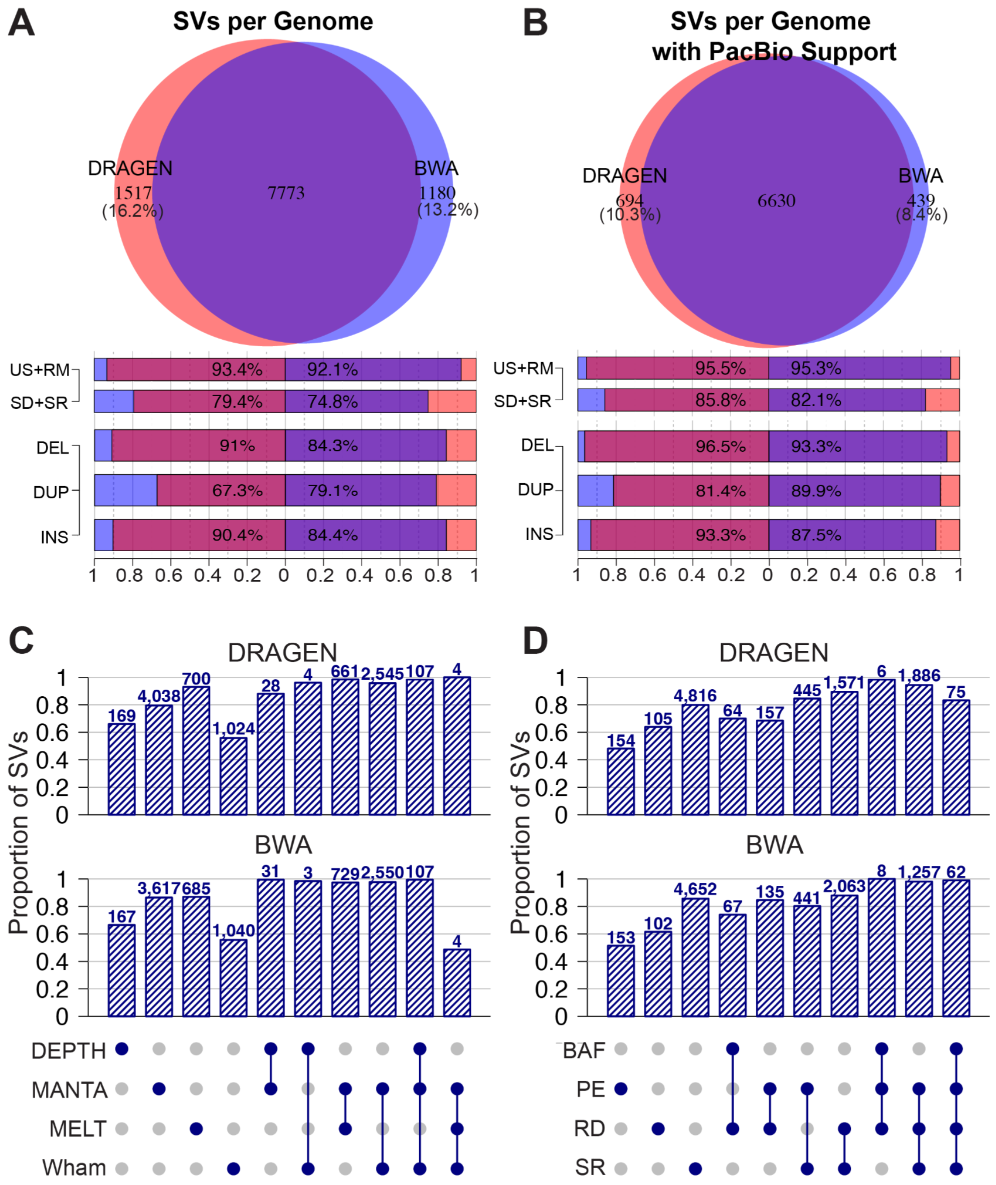

Figure 8. Characteristics of the overlap between DRAGEN and BWA SVs. A-B. Overlap between BWA and DRAGEN callsets of A) SV genotypes in the same genome and B) SV genotypes in the same genome that are supported by lrWGS data, and the breakdown of overlap by different genomic context and SV type. C-D. Proportion of DRAGEN and BWA SVs in each individual genome that are overlapped by the other aligner broken down by C) GATK-SV component algorithms and D) alignment evidence.

Discussion

Previous studies have indicated that the srWGS alignments are usually confounded by the complexity of genomic sequences, so SV discovery in the complex SD and SR sequences is more challenging and prone to higher false positive rates [5,25]. The lower concordance between DRAGEN and BWA SVs in SD and SR regions indicate a positive correlation between the concordance of the aligners and the quality of SVs, which is further supported by the observations that SVs with PacBio support shared higher concordance between aligners. It should be noted that the SVs included in this comparison were the direct output from GATK-SV pipeline, which maximizes sensitivity, and downstream filtering and refinements to improve precision were not applied in order to preserve the comparability of the BWA and DRAGEN callsets. Therefore, the discordance between aligners is likely a result of both technical alignment differences and false positive SVs. In our next round of benchmarking, we plan to evaluate the DRAGEN 3.7.8 aligner, enlarge the sample set to include all 3,202 1KGP samples, and perform SV filtering and refinement.

References

[1] "All of Us Genomic Quality Report" *All of Us Research Program*, https://support.researchallofus.org/hc/en-us/articles/4617899955092-All-of-Us-Beta-Release-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] 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

[5] 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

[6] 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

[7] Fu, Jack M et al. "Rare coding variation provides insight into the genetic architecture and phenotypic context of autism." *Nature genetics* vol. 54,9 (2022): 1320-1331. doi:10.1038/s41588-022-01104-0

[8] 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.

[9] P. Ebert, P. A. Audano, Q. Zhu et al., Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021).
[10] **PacBio structural variant calling and analysis tools (PBSV)**, Retrieved March 3, 2023, from https://github.com/PacificBiosciences/pbsv.
[11] 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.
[12] GATK Team "SVConcordance (Beta) – GATK." *GATK*, 20 Mar. 2023, https://gatk.broadinstitute.org/hc/en-us/articles/13832773767963-SVConcordance-BETA-.
[13] 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
[14] 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
[15] 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.
[16] 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
[17] 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
[18] Li, Heng, and Richard Durbin. "Fast and accurate short read alignment with Burrows-Wheeler transform." *Bioinformatics (Oxford, England)* vol. 25,14 (2009): 1754-60. doi:10.1093/bioinformatics/btp324
[19] Karczewski, Konrad J et al. "The mutational constraint spectrum quantified from variation in 141,456 humans." *Nature* vol. 581,7809 (2020): 434-443. doi:10.1038/s41586-020-2308-7
[20] Taliun, Daniel et al. "Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program." *Nature* vol. 590,7845 (2021): 290-299. doi:10.1038/s41586-021-03205-y
[21] Caetano-Anolles, Derek. "Introducing DRAGMAP, the New Genome Mapper in DRAGEN-GATK." *GATK*, gatk.broadinstitute.org/hc/en-us/articles/4410953761563-Introducing-DRAGMAP-the-new-genome-mapper-in-DRAGEN-GATK.
[22] Kosugi, Shunichi et al. "Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing." *Genome biology* vol. 20,1 117. 3 Jun. 2019, doi:10.1186/s13059-019-1720-5
[23] Tattini, Lorenzo et al. "Detection of Genomic Structural Variants from Next-Generation Sequencing Data." *Frontiers in bioengineering and biotechnology* vol. 3 92. 25 Jun. 2015, doi:10.3389/fbioe.2015.00092
[24] Collins, Ryan L et al. "Defining the diverse spectrum of inversions, complex structural variation, and chromothripsis in the morbid human genome." *Genome biology* vol. 18,1 36. 6 Mar. 2017, doi:10.1186/s13059-017-1158-6

[25] Cameron, Daniel L et al. “Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software.” *Nature communications* vol. 10,1 3240. 19 Jul. 2019, doi:10.1038/s41467-019-11146-4

Appendix A: Additional technical details for the DRAGEN 3.4.12 evaluation

We compared the performance of the BWA-MEM 0.7.15 (referred to as BWA) and DRAGEN 3.4.12 (referred to as DRAGEN) aligners for detecting structural variants using the GATK SV pipeline. The experimental setup involved preparing the raw data, creating Terra workspaces for each dataset, running the GATK-SV variant calling pipeline for each dataset, and analyzing the resulting variant calls. This setup enabled us to compare the performance of these two aligners and assess any differences in the variant calling results.

Sample Selection

161 samples were selected from the 3,202 NYGC high-coverage 1000 Genomes samples [3]. 23 samples were selected because they had matched PacBio lrWGS data [9]. The remaining 138 samples are part of trios, including 8 trios containing at least one sample with lrWGS data. Selecting samples with either matched lrWGS data or close family members provided options for validating SV calls by comparisons to lrWGS data or by examining Mendelian violation rates.

Realignment with DRAGEN 3.4.12

The publicly available BWA-aligned CRAM files, CRAM index files, and gVCF files were used for the BWA inputs [3]. To generate the DRAGEN inputs, the BWA-aligned CRAM files were first sorted and converted to FASTQ. Then DRAGEN 3.4.12 was used to align the FASTQ and produce a gVCF, following the *All of Us* DRAGEN 3.4.12 GRCh38 specifications. This process was performed in Amazon Web Services (AWS) using AWS Batch. Table A.1 shows these steps in more detail.

Table A.1 -- Steps to realign 1KGP CRAMs with DRAGEN 3.4.12

Step	Tool	Version Used	Command
Sort CRAM files	Samtools	1.12	samtools sort -n --reference GRCh38_full_analysis_set_plus_decoy_hla.fa -o {output} {input}
Convert CRAM to FASTQ	Samtools	1.12	samtools fastq --reference GRCh38_full_analysis_set_plus_decoy_hla.fa -1 {output_r1} -2 {output_r2} {input}

Compress FASTQ	gzip	1.3.12	gzip {input}
Produce DRAGEN 3.4.12 CRAM and gVCF	DRAGEN	3.4.12	parameters are identical to the *All of Us* DRAGEN 3.4.12 GRCh38 specifications

Callset generation with GATK-SV

The GATK-SV pipeline for structural variant calling was applied to the 161 samples in parallel from the BWA-aligned inputs and the DRAGEN-aligned inputs. The exact same steps were followed and the exact same version of the GATK-SV code was used for each set of inputs so that the only difference between the production of the two callsets was the aligner.

Table A.2 shows the steps of the GATK-SV pipeline that were applied for this analysis, available in the GATK-SV GitHub repository (https://github.com/broadinstitute/gatk-sv). The table also lists the GitHub release version that was used for each of the steps. The versions vary between steps because the latest version of each workflow available at the time was used. Some of the workflow names and versions differ from those used for the v7 srWGS SV release [1] because this analysis was completed earlier than the main callset.

Table A.2-- GATK-SV Pipeline Versions Used for Evaluating DRAGEN against BWA

Workflow/Step Name	Version Used
GatherSampleEvidence	v0.21-beta
EvidenceQC	v0.21-beta
TrainGCNV	v0.23-beta
GatherBatchEvidence	v0.23-beta
ClusterBatch	v0.21-beta
GenerateBatchMetrics	v0.21-beta
FilterBatchSites	v0.21-beta
PlotSVCountsPerSample	v0.21-beta
FilterBatchSamples	v0.21-beta
MergeBatchSites	v0.21-beta
GenotypeBatch	v0.21-beta
RegenotypeCNVs	v0.21-beta
MakeCohortVcf	v0.22-beta
AnnotateVcf	v0.23-beta