

Size dependent loss of structural variant allele counts in pangenome derived VCFs

Manjunath M

iammanjunath59@gmail.com

University of Mysore

Raviraj V S

Central Silk Board

Shakunthala V

University of Mysore

Research Article

Keywords: Pangenome graph, Structural variation, Allele frequency, Variant representation, VCF, Bubble decomposition, Segmental duplication

Posted Date: October 29th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11373989/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background.

Pangenome graphs represent structural variation as divergences between haplotype paths, but almost all downstream analysis consumes VCF, so graphs are projected onto a linear reference by a standard conversion chain. Whether structural variant allele counts survive that projection is unknown, because real data provides no ground truth.

Results.

We simulated haplotype panels on real GRCh38 sequence with known carriers for every structural variant, giving 4,056 variants across two chromosomes. Recovery of the true allele count falls with variant size, from 0.988 at 100 bp to 0.553 at 30 kb, with near identical coefficients on $\log_{10}$ size on the discovery and held out chromosomes (-2.173 and -2.183). Recovery declines further with segmental duplication content, to 0.375 for 30 kb variants. The loss arises in bubble decomposition rather than post-processing: `vcfbub` changes nothing, and `vcfwave` raises readable recovery from 0.466 to 0.865 overall yet only to 0.553 at 30 kb. Graphs built from identical panels by PGGB and Minigraph-Cactus behave alike. Computing allele counts from haplotype path traversals instead recovers 0.992 overall and 0.986 at 30 kb, independent of size. In the HPRC Release 2 pangenome, a 62 kb deletion carried by 216 of 461 haplotypes appears as 84 alleles whose largest count is 25.

Conclusions.

Allele counts of large structural variants, especially in segmental duplications, are fragmented when pangenome graphs are decomposed into VCF records. The effect belongs to the representation, not to any one graph builder, and counting alleles from haplotype path traversals recovers what the conversion loses.

Background

Structural variants alter more base pairs per genome than all single nucleotide changes combined, and they are disproportionately represented among variants of large phenotypic effect (Collins et al. 2020). They have nevertheless remained harder to study than point mutations, largely because they are harder to represent. A single nucleotide change is fully described by a position and a substituted base. A structural variant is described by two breakpoints whose placement is often ambiguous, by sequence that may be absent from the reference entirely, and by a relationship to nearby variants that a position based format is not designed to express (Hickey et al. 2020; Ebert et al. 2021).

Pangenome graphs were developed in part to address this. By representing many assembled haplotypes in one structure, a graph encodes each haplotype as a path and each variant as a divergence between paths, so that sequence absent from any single reference is nevertheless present in the model (Garrison et al. 2018). The Human Pangenome Reference Consortium has now released two such resources built from phased diploid assemblies (Liao et al. 2023; Lucas et al. 2026), constructed by two independent pipelines, PGGB (Garrison et al. 2024) and Minigraph-Cactus (Hickey et al. 2024). Within the graph, a structural variant is not an annotation but a topological feature: the set of haplotype paths that diverge from the reference and rejoin it.

Almost no downstream analysis consumes the graph in that form. Association testing, population genetic inference, clinical interpretation and variant annotation all read VCF, so graph based resources are distributed as VCF alongside the graph itself, and the projection from one to the other is performed by a standard pipeline: *vg deconstruct* enumerates the alleles of each snarl (Garrison et al. 2018; Paten et al. 2018), *vcfbub* filters and flattens nested bubbles (Garrison 2022), and *vcfwave* realigns complex records into simpler ones (Garrison et al. 2022). This chain is not an implementation detail. It is the interface through which essentially every user of a pangenome encounters its structural variation.

The projection is not straightforward, because the two representations do not share a notion of what an allele is. In a graph, an allele is a traversal: the ordered sequence of segments a haplotype follows between two points where all haplotypes agree. In a VCF, an allele is a reference span paired with a replacement string at a single coordinate. Traversals nest, overlap and differ at many scales at once, whereas VCF records are flat and positional. Mapping the first onto the second requires decisions about where events begin and end and about which distinctions between traversals are worth preserving, and different decisions yield different callsets from the same graph.

Several lines of evidence indicate that these decisions have consequences. Breakpoint placement for structural variants is systematically biased by nearby small polymorphisms, affecting roughly 5% of the structural variants called in a human genome, and graph based normalisation, although it makes calls consistent across samples, introduces technical biases of its own (Audano and Beck 2024). Genotyping accuracy is higher when graphs are built directly from assembly alignments than when they are built from an intermediate VCF catalogue of the same variants, which indicates that passing through the VCF format loses information the graph contained (Hickey et al. 2020). The Human Pangenome Reference Consortium itself notes that segmental duplication regions in its second release contain a larger biallelic fraction and relatively fewer highly multiallelic sites than the first, and attributes this in part to reduced spurious splitting of similar alleles in complex duplicated regions (Lucas et al. 2026).

What has not been measured is whether allele counts survive the projection. This is the quantity that downstream analysis actually consumes. Population frequency estimates for 167,000 structural variants across 5,202 genomes, and the expression quantitative trait loci identified from them, were derived from graph based genotypes (Sirén et al. 2021). In tomato, structural variants genotyped against a graph pangenome recovered a substantial fraction of previously missing heritability, with association tests

conducted on allele frequencies obtained from the graph (Zhou et al. 2022). Constraint, selection and disease association arguments all rest on how often an allele is seen, not merely on whether it is seen. If the count attached to a structural variant in a pangenome derived VCF differs from the number of haplotypes that actually carry it in the graph, every such inference inherits the discrepancy, and does so silently, because nothing in the VCF records that a count has been split or merged.

The difficulty in assessing this on real data is that there is no ground truth. When a graph and its derived VCF disagree about how many haplotypes carry an event, no external source establishes which is right, and in complex regions it is not always clear that the question is well posed, since the two representations may be partitioning the same sequence differences into different events. Comparisons between real callsets can therefore demonstrate discordance but cannot quantify error.

We address this by constructing benchmarks in which the answer is known. Haplotype panels were simulated on real GRCh38 sequence, with ancestry generated under the coalescent (Baumdicker et al. 2022) and structural variants placed at known positions in known carriers, so that the true allele count of every variant is fixed by construction while the sequence context retains the repeat architecture that makes graph construction difficult. Because the simulator is deterministic, the same panel can be built into graphs by different tools, which makes the comparison between builders exact rather than approximate.

Using 4,056 structural variants across two chromosomes, we ask three questions. First, how much of the true allele count survives the conversion of a pangenome graph to VCF, and does the loss depend on variant size and on sequence context? Second, where in the conversion chain does any loss occur, and is it a property of the graph builder or of the VCF representation itself? Third, can allele counts be recovered by reasoning over haplotype path traversals directly, without an intermediate VCF? We then ask whether the mechanism identified in simulation is present in the published human pangenome, comparing two representations distributed with the same release against each other.

We find that recovery declines steeply with variant size, that the decline originates in bubble decomposition rather than in post-processing, that it appears equally in graphs built by two independent pipelines, and that counting alleles from traversals recovers what the VCF route loses. In the HPRC Release 2 pangenome, a deletion of approximately 62 kb carried by 216 of 461 haplotypes is represented in the raw bubble decomposition by 84 distinct alleles whose largest count is 25, and in the released callset by four separate records that partition its carriers.

Results

A benchmark in which every allele count is known exactly

Whether a pangenome VCF preserves the allele count of a structural variant cannot be assessed on real data, because no real dataset supplies the true count. We therefore built a benchmark in which the truth is known by construction. Haplotype panels were simulated on real GRCh38 sequence, with ancestry

generated under the coalescent using msprime (Baumdicker et al. 2022) and structural variants placed at known positions, in known carriers, at known frequencies. Because the simulator is deterministic given a seed, the same panel can be handed to different graph builders, which makes builder comparison exact rather than approximate.

The primary sweep covers 54 windows of one megabase on chromosome 20 at two seeds each: 108 replicates, 24 structural variants per replicate, 2,592 variants in total. Variants are stratified across six size classes (100 bp, 300 bp, 1 kb, 3 kb, 10 kb and 30 kb, with 432 variants per class), so that size effects are estimated on balanced data rather than on whatever the frequency spectrum happens to supply. The set comprises 1,886 deletions, 685 duplications and 21 inversions. A held out set of 61 replicates on chromosome 16 (1,464 variants) was analysed only after the chromosome 20 analysis was complete, giving 4,056 variants with exact ground truth in total.

Graphs were built with PGGB (Garrison et al. 2024) as the primary builder and converted to VCF using the pipeline established by the Human Pangenome Reference Consortium: vg deconstruct (Garrison et al. 2018) over all snarls including nested ones (Paten et al. 2018), followed by vcfbub (Garrison 2022) and vcfwave (Garrison et al. 2022). A variant was scored as recovered when its true allele count appears among the alternate allele counts at an overlapping site, with overlap defined by a tolerance of $\max(200 \text{ bp}, \text{half the variant length})$. This criterion is permissive by design: it credits a site even when the correct count sits among many alternates. Where the additional requirement that the site be biallelic changes the conclusion, both are reported, and the difference between the two turns out to be the most informative quantity in the analysis.

The design separates the two covariates of interest. Variance inflation factors for centred $\log_{10}$ variant size and for segmental duplication content are both 1.0000, and across windows the correlation between mean $\log_{10}$ variant size and segmental duplication content is -0.101 (Spearman $\rho = -0.238$ on untransformed size). GC content ranges from 0.370 to 0.492 across windows, correlates only weakly with segmental duplication content ($r = +0.253$), and is not a confounder: adding it to the model leaves it non significant ($p = 0.60$) and moves the segmental duplication coefficient only from -0.374 to -0.353 . No simulated variant falls below the 50 bp minimum segment length used by the graph route (0 of 2,592), so that threshold removes nothing from the benchmark.

Allele count recovery from VCF declines steeply with variant size

Across the 2,592 chromosome 20 variants, the standard conversion route recovered the true allele count for 86.5 % of variants. That aggregate conceals a strong dependence on size (Fig. 1a, Table 1).

SV size class	Variants (n)	Recovered
100 bp	432	0.988
300 bp	432	0.977
1 kb	432	0.988
3 kb	432	0.898
10 kb	432	0.782
30 kb	432	0.553

Table 1. Recovery of the true allele count by variant size class, PGGB graphs, standard conversion route, 108 replicates on chromosome 20.

Recovery is essentially complete up to 1 kb and then falls away, reaching 0.553 in the 30 kb class: nearly half of the largest variants have no recoverable allele count anywhere in the converted VCF. In a binomial generalised linear model of recovery on centred $\log_{10}$ variant size and segmental duplication content, with standard errors clustered by window, the size coefficient is -2.173 (SE 0.139, 95 % CI -2.446 to -1.900 , $z = -15.6$, $p = 9.6 \times 10^{-55}$), an odds ratio of 0.114 per tenfold increase in variant size. Clustering matters: the 24 variants within a replicate share a genomic context and a graph build, and the two seeds at a window share the window, so naive standard errors are anticonservative, particularly for the segmental duplication term, which is identified between windows rather than within them. Full coefficient tables for both chromosomes are given in Additional file 1: Table S4.

The held out chromosome 16 replicates reproduce the size effect almost exactly: -2.183 (SE 0.252, 95 % CI -2.676 to -1.690 , $p = 4.1 \times 10^{-18}$) across 61 replicates and 1,464 variants, against -2.173 on chromosome 20. The two point estimates differ by 0.010, well inside either interval. The size effect is therefore not a property of one chromosome's sequence composition.

Recovery also declines with the segmental duplication content of the surrounding window: -0.419 per ten percentage points of segmental duplication on chromosome 20 (SE 0.063, 95 % CI -0.544 to -0.295 , $p = 3.6 \times 10^{-11}$) and -0.281 on chromosome 16 (SE 0.061, $p = 3.7 \times 10^{-6}$). The two effects are additive rather than interacting; adding a size by segdup term leaves the interaction indistinguishable from zero (-0.123 , SE 0.206, $p = 0.55$). Stratifying by segmental duplication content shows the two effects compounding at the largest sizes (Fig. 1b ; Additional file 1: Table S3). On chromosome 20, recovery of 30 kb variants is 0.571 in low segdup windows and 0.375 in high segdup windows, while at 100 bp the two strata are indistinguishable (0.987 and 1.000). Chromosome 16 shows the same pattern across a wider segdup range, from 0.994 against 0.956 at 100 bp to 0.562 against 0.485 at 30 kb.

The loss originates in bubble decomposition, not in post-processing

vcfbub and vcfwave exist specifically to resolve nested bubble structure, so their contribution should be measured rather than assumed. We regenerated the raw vg deconstruct output and the intermediate produced after vcfbub for all 108 replicates and evaluated all three stages under the same criterion (Additional file 1: Fig. S1).

Under the permissive criterion the three stages are nearly indistinguishable: 0.878 for the raw deconstruction, 0.877 after vcfbub and 0.865 after vcfwave (Additional file 1: Table S1) . The correct allele count is already present in the raw output for the large majority of variants, and post-processing does not add to it. In the largest size class vcfwave in fact reduces recovery slightly, from 0.623 to 0.553. Requiring in addition that the site be biallelic, so that the correct count is identifiable without prior knowledge of the answer, separates the stages sharply (Table 2).

SV size class	Raw	After vcfbub	After vcfwave
100 bp	0.975	0.970	0.988
300 bp	0.903	0.903	0.977
1 kb	0.597	0.595	0.988
3 kb	0.227	0.227	0.898
10 kb	0.062	0.062	0.782
30 kb	0.035	0.035	0.553
Overall	0.466	0.465	0.865

Table 2. Readable recovery, defined as the true allele count present at an overlapping site that carries at most two alleles, at three points in the conversion chain.

Two conclusions follow. First, vcfbub is inert with respect to allele count recovery at these parameters: it changes the overall rate by 0.001 and leaves every size class unchanged to three decimal places. Second, in the raw decomposition the allele counts of large variants are present but not readable. For 30 kb variants the true count appears somewhere at the site in 62 % of cases but in a usable form in only 3.5 %. Large variants in the raw output are represented as sites carrying many alternate alleles, among which the true count is one value among many. vcfwave resolves these sites into biallelic records and thereby raises readable recovery from 0.466 to 0.865 overall, and from 0.035 to 0.553 at 30 kb. It repairs most of the problem and it does not repair all of it. The fragmentation of the largest variants originates upstream, in how bubble decomposition enumerates alleles, and post-processing that operates on the already decomposed VCF cannot fully reconstruct distinctions the decomposition did not make.

Counting alleles from graph traversals avoids the problem

If the loss arises in converting graph topology into allele records, then computing allele counts directly from haplotype path traversals should avoid it. graphsv walks the reference path, identifies the segments

at which haplotype paths diverge from and rejoin it, and counts the haplotypes taking each traversal, without producing an intermediate VCF. On the same 2,592 chromosome 20 variants, graphsv recovered the correct allele count for 99.2 %, and its recovery is nearly independent of variant size (Fig. 2, Table 3).

SV size class	Graph route	VCF route
100 bp	0.972	0.988
300 bp	0.998	0.977
1 kb	1.000	0.988
3 kb	1.000	0.898
10 kb	0.998	0.782
30 kb	0.986	0.553
Overall	0.992	0.865

Table 3. Allele count recovery by route, PGGB graphs, 2,592 variants across 108 replicates on chromosome 20.

The two routes agree for small variants and diverge as size increases, reaching a difference of 0.433 in the 30 kb class. The only class in which the graph route is weaker is 100 bp, at 0.972. These twelve failures are not miscounts: in every case graphsv called no site at the locus and returned an empty allele set rather than an incorrect allele count. The true lengths of those variants range from 52 bp to 182 bp, all above the 50 bp minimum segment length, so the shortfall reflects detection sensitivity near that floor rather than variants being excluded by it.

The graph route is also markedly faster. On a one megabase window with 20 haplotypes, graphsv completed in 0.08 s to 0.13 s across repeated runs with a peak resident set of 27 MB, against 6.07 s to 6.13 s and 141 MB for vg deconstruct followed by vcfbub and vcfwave on the same graph: roughly a fiftyfold difference in wall time while producing the more accurate counts.

The effect is a property of VCF representation, not of one graph builder

Because the simulator is deterministic, identical haplotype panels can be built into graphs by different tools. We rebuilt all 54 chromosome 20 windows at seed 1 with Minigraph-Cactus (Hickey et al. 2024), asserting panel identity against the PGGB run before each build; all 54 truth tables were byte identical between the two arms, so any difference is attributable to the builder alone. Minigraph-Cactus applies vcfbub internally, so only vcfwave was added to its output, matching the PGGB pipeline (Table 4).

Builder	Variants (n)	VCF route	Graph route
PGGB	2,592	0.865	0.992
Minigraph-Cactus	1,296	0.875	0.988

Table 4. Allele count recovery by builder and route, on byte identical haplotype panels.

The builders differ by about one percentage point on each route, and the size dependence is the same in both. The Minigraph-Cactus VCF route falls from 0.991 at 100 bp to 0.569 at 30 kb, close to PGGB's 0.553 in the same class, while its graph route stays between 0.968 and 0.995 across all size classes (Additional file 1: Table S2) . Two independently developed alignment and graph construction pipelines therefore produce the same pattern, which localises the loss to what they share, namely the decomposition of graph topology into allele records, rather than to graph construction.

Robustness

Segmental duplication effect

The segdup association is carried by a small number of windows at the top of the covariate range, so we tested it three ways. Permuting segdup labels across windows, holding the size effect and all within window structure fixed (2,000 permutations), places the observed coefficient of -0.419 far outside the null distribution (null mean $+0.013$, SD 0.179 , 2.5th to 97.5th percentiles -0.304 to $+0.396$; one sided $p = 5.0 \times 10^{-4}$). Dropping the k highest segdup windows in turn leaves the coefficient negative throughout and in fact more strongly negative as the covariate range narrows (-0.461 at $k = 1$, -1.186 at $k = 3$, -2.224 at $k = 5$), with standard errors inflating as the upper range is removed. Leave one window out shifts the coefficient by at most 0.042 (Additional file 1: Table S5). The association is therefore not an artefact of any single window, although its magnitude is estimated from a covariate range whose upper end is sparsely sampled: 40 variants per size class in the high segdup stratum against 392 in the low segdup stratum on chromosome 20.

Merge criterion

graphsv merges adjacent divergent segments that share a carrier set into a single call, which risks fusing genuinely distinct variants. Auditing the 2,512 calls made across the 108 replicates against the known truth, 2,128 (84.7 %) contain exactly one true variant, 52 (2.07 %) span two or more, and 332 (13.2 %) contain none. In the reverse direction, 2,570 of 2,592 true variants (99.2 %) are represented by exactly one call, 2 by more than one (residual split rate 0.0008) and 20 by none. The criterion had genuine opportunity to err: 79 pairs of true variants share an identical carrier set, and 63 of those lie within 50 kb of one another, so complete linkage at short range is common in this design rather than absent (Additional file 1: Table S6). The fusions that do occur follow that pattern. A 22,335 bp and a 1,923 bp variant, both at allele count 9, are written as a single 31 kb call; a 30,000 bp and a 120 bp variant, both at allele count 3, likewise. The surviving allele count is correct in such cases, which is why the recovery

criterion scores them as successes, but the variant count is not: two events are reported as one, at the size of their span.

Variant placement scheme

Recovery under realistic rather than stratified placement is substantially higher. Across 108 replicates in which variant sizes are drawn from an empirical frequency spectrum rather than balanced across size classes, mean recovery is 0.985. Large variants are rare under realistic sampling, so the aggregate is dominated by the size classes in which conversion performs well. This does not weaken the size effect, which is the same effect weighted by how often large variants occur, but it does bound the practical impact: analyses dominated by small variants will be little affected, whereas analyses that depend on large structural variants will be affected substantially.

The mechanism is present in the published human pangenome

Simulation supplies ground truth; real data supplies the demonstration that the mechanism occurs in a graph that is in active use. We queried the HPRC Release 2 Minigraph-Cactus pangenome (Lucas et al. 2026), version 2.1, directly from its public S3 location, comparing the raw bubble decomposition against the released post-processed VCF. The partition treated as the reference here is vg's own, recorded in the AT traversal field of the raw callset; graphsv is not applied to these data at any point. The comparison is between two representations produced by the same published pipeline, and is therefore a demonstration that the two disagree rather than a test of either against truth.

At chr20:25,783,205 a top level bubble spans 68,466 bp of GRCh38 and carries 328 alternate alleles across 461 haplotypes (Fig. 3). Reconstructing traversals from the AT field shows that 84 of those alleles share a single pair of breakpoint nodes and a gap of 764 reference path nodes out of 1,177: they are one deletion of approximately 62 kb, differing only in small indels within the sequence they retain. That deletion is carried by 216 of 461 haplotypes, a frequency of 0.469. The largest single allele count reported at the site is 25, a frequency of 0.054, an understatement of 8.6 fold.

This is not a consequence of post-processing. It is present in the raw decomposition, where allele identity is the full traversal sequence across the bubble, so haplotypes that differ anywhere within 68 kb are assigned distinct alleles. At bubbles of this span the allele count vector is closer to a haplotype spectrum than to a variant frequency spectrum. The consortium reports a related observation from the opposite direction: in segmental duplication regions, Release 2 shows a larger biallelic contribution and relatively fewer sites with more than three alleles than Release 1 (Liao et al. 2023), which the authors attribute in part to reduced spurious splitting of similar alleles in complex duplicated regions (Lucas et al. 2026).

In the released VCF, vcfwave recovers the event. Four records carry it, holding 200, 12, 3 and 1 haplotype copies, summing exactly to 216, and no haplotype copy appears in more than one record: the carriers are partitioned cleanly rather than duplicated or lost. The dominant record reports an allele frequency of

0.434 against the true 0.469. The frequency is therefore approximately correct while the event is reported as four segregating variants rather than one, an error of variant counting rather than of frequency estimation.

Extending this to all large deletions on the chromosome, a scan of the raw VCF identified 24 deletion events of at least 20 kb (median 57,187 bp, range 20,335 bp to 1,038,174 bp; 21 top level and 3 nested). Classifying the 41 distinct deletion events resolved at those bubbles gives 10 recovered by a single matching record, 4 fragmented across records, 2 represented at a different scale, 1 partial and 24 with no size matching record (Additional file 1: Table S8). The last category requires care: all 24 occur in only four bubbles, every one of which has a reduced allele number (AN 49, 209 and 363, against 460 to 463 elsewhere), and 23 of the 24 fall in just three of them. In two of those bubbles the median non carrier haplotype carries as much deleted sequence as the median carrier, so the comparison has no contrast and is degenerate. Restricted to well covered bubbles, most large deletions are recovered cleanly and a minority fragment (Additional file 1: Table S7) .

Two further cases show why apparent absence must be checked rather than assumed. At chr20:25,936,878 a 26 kb deletion at frequency 0.241 has no size matching record, yet each of its carriers carries 47,145 bp of deleted sequence in the released VCF: the event is represented at the scale of the enclosing bubble rather than at the scale of the event. At chr20:25,984,965 a 4.9 kb deletion at frequency 0.200 sits inside a larger deleted region. In neither case is sequence lost; in both, the two representations disagree about the unit of variation.

That disagreement has no resolution in real data, because no ground truth exists for where one structural variant ends and another begins. It is precisely the limitation that the simulation arm is designed to remove, and it is the reason the quantitative claims in this work rest on simulated panels with known truth while the HPRC analysis serves to demonstrate that the mechanism is present in a resource in current use.

Discussion

Allele counts of structural variants are not reliably preserved when a pangenome graph is converted to VCF, and the loss is systematic rather than sporadic. Recovery falls from 0.988 for 100 bp variants to 0.553 for 30 kb variants, an odds ratio of 0.114 per tenfold increase in size, and the estimate replicates on a held out chromosome to within 0.010. It falls further with segmental duplication content, so that 30 kb variants in high segdup windows are recovered at 0.375 against 0.571 elsewhere. Neither effect is an artefact of a particular graph builder: PGGB and Minigraph-Cactus, given byte identical haplotype panels, differ by about one percentage point on each route and show the same size dependence. Computing allele counts from haplotype path traversals instead recovers 0.992 overall and 0.986 in the largest class, at roughly a fiftieth of the wall time.

Why decomposition loses counts

A single mechanism accounts for both the size effect and the segmental duplication effect. In a bubble decomposition, an allele is the entire traversal between the points at which all haplotypes agree, so two haplotypes share an allele only if they are identical across the whole span. The probability that they are identical declines with the length of the span and with the density of variation within it. A large structural variant is bracketed by distant anchors and therefore encloses many smaller polymorphisms, each of which is sufficient to split its carriers into separate alleles. The event itself is faithfully encoded in the graph topology, but the allele level summary counts traversals rather than events.

Segmental duplications compound this because they carry more variation per base than unique sequence. Mutation rates and interlocus gene conversion are both elevated in human segmental duplications (Vollger et al. 2023), so a bubble of a given span in such a region encloses more polymorphism and fragments into more traversals. The two covariates act additively in our data, with no detectable interaction, which is what this account predicts: span and per base diversity enter multiplicatively into the probability of traversal identity and therefore additively on the log odds scale.

This explains the pattern observed in the raw deconstruction, where the correct allele count is present at an overlapping site for 62% of 30 kb variants but readable at a biallelic site for only 3.5%. The information is not destroyed; it is distributed across alternates in a way that cannot be resolved without knowing the answer. vcfwave addresses exactly this by realigning complex records into simpler ones, and it is effective, raising readable recovery from 0.466 to 0.865 overall. What it cannot do is reconstruct a distinction the decomposition never made. Where two carriers of one event were assigned different alleles because they differ elsewhere in the bubble, realignment can separate the shared deletion from the incidental differences only when the alignment is unambiguous, and at 30 kb it succeeds for 0.553 of variants rather than all of them.

Relation to existing work on variant representation

Ambiguity in how a variant is written is a long standing problem in the VCF format. Normalisation by left alignment and parsimony gives a unique representation for an isolated variant (Tan et al. 2015), but the authors of that work already noted that adjacent or complex variants admit no such unique form. A bubble in a pangenome graph is precisely the complex case: several variants of different scales, occupying the same interval, observed in combination across haplotypes.

Two recent findings bear directly on ours. Audano and Beck (2024) showed that structural variant breakpoints are systematically displaced by nearby small polymorphisms, affecting roughly 5% of the structural variants in a human genome, and that graph based normalisation, while it makes calls consistent across samples, introduces technical biases of its own. Their analysis concerns where an event is placed; ours concerns how many haplotypes it is recorded as having, and the two are consequences of the same underlying difficulty, namely that small variation inside a large event perturbs how that event is described. Hickey et al. (2020) found that graphs built directly from assembly alignments genotype structural variants better than graphs built from an intermediate VCF catalogue of the same variants, which indicates that the VCF representation loses information present in the

alignments. Our result is the allele count analogue of that observation, measured against known truth rather than inferred from downstream accuracy.

The Human Pangenome Reference Consortium reports a related observation from its own data: segmental duplication regions in Release 2 contain a larger biallelic fraction and relatively fewer highly multiallelic sites than in Release 1, which the authors attribute in part to reduced spurious splitting of similar alleles in complex duplicated regions (Lucas et al. 2026). Improvements in assembly and alignment quality therefore mitigate the problem, which is consistent with our account: better alignment reduces the number of spurious distinctions between traversals. It does not remove the problem, because genuine variation inside a large bubble splits traversals just as effectively as alignment error does.

Consequences for analyses that use pangenome derived allele frequencies

Allele frequency is the quantity most downstream analysis consumes. Population frequency estimates for 167,000 structural variants across 5,202 genomes, and the expression quantitative trait loci derived from them, rest on graph based genotypes (Sirén et al. 2021); structural variants genotyped against a graph pangenome recovered a substantial part of previously missing heritability in tomato, with association testing performed on graph derived frequencies (Zhou et al. 2022); and constraint and selection arguments for structural variation are framed in terms of the frequency spectrum (Collins et al. 2020). An allele count that is understated by a factor of eight, as at the HPRC locus examined here, propagates into all of these without leaving any trace in the VCF that it has done so.

Two distinct kinds of error need separating, because they have different consequences and different magnitudes. At chr20:25,783,205 in the HPRC Release 2 callset, the dominant record reports an allele frequency of 0.434 against a true 0.469, an error of about 7%, which for most purposes is tolerable. The same event, however, is written as four separate records whose carrier sets partition its 216 haplotype copies without overlap. Any analysis that counts segregating structural variants, computes structural variant load per genome, or measures structural variant density along a chromosome overcounts fourfold at this locus, and nothing in the records indicates that they belong together. Frequency estimation is comparatively robust; variant counting is not. We would encourage explicit reporting of which of the two a given analysis depends on.

The practical scope of the problem is bounded by how often large structural variants occur. Under realistic rather than stratified placement, where variant sizes follow an empirical frequency spectrum, mean recovery is 0.985, because large variants are rare and the aggregate is dominated by the size classes that convert cleanly. Analyses dominated by small structural variants are therefore largely unaffected. The exposure is concentrated in large events and in segmental duplications, and those are not incidental categories: duplicated regions are enriched for genes, are catalysts for recurrent copy number variation, and are disproportionately implicated in disease (Vollger et al. 2023). The regions where the representation is least reliable are among those where accurate frequencies matter most.

Recommendations

Three changes would reduce the exposure without requiring new infrastructure. First, allele counts intended for population genetic use can be computed on the graph rather than read from the projected VCF; graphsv does this in a fraction of the time the conversion itself takes, and the approach is not specific to our implementation. Second, released pangenome VCFs can retain the traversal field. The HPRC Release 2 wave callset has AT removed by a bcftools annotate step in the release pipeline, and with it the information needed to recover the graph partition downstream; the raw callset retains it, which is what made the analysis reported here possible at all. Third, analyses that count structural variants rather than estimate their frequencies should be performed against the graph or against a callset in which records belonging to one event are grouped, since nothing in a flat VCF marks that relationship.

More generally, some analyses already bypass the VCF intermediate. Genotyping methods that operate on bubbles and haplotype paths directly, whether by read mapping to the graph (Hickey et al. 2020; Sirén et al. 2021) or by k-mer inference over haplotype resolved panels (Ebler et al. 2022), do not inherit the decomposition step at all. The results here suggest that frequency estimation belongs in the same category.

Limitations

The benchmark uses 20 haplotypes per panel, an order of magnitude fewer than the current human pangenome. Fewer haplotypes means fewer distinct traversals per bubble, so the simulation is likely to understate rather than overstate the fragmentation that occurs in a large panel; the direction of this bias is conservative but its magnitude is not established here.

Simulated structural variants are constrained to be non overlapping and separated by at least 200 bp, so each corresponds to a single bubble. The benchmark therefore does not test overlapping or nested structural variation, which is common in real genomes and which the HPRC analysis suggests is where representations diverge most sharply. The composition is also uneven: 1,886 deletions, 685 duplications and 21 inversions, with inversions synthesised rather than drawn from the empirical callset because short read derived catalogues undercall them (Collins et al. 2020; Ebert et al. 2021). Inversion recovery measured on randomly placed inversions lacking flanking inverted repeats should be read as an upper bound.

The largest simulated class is centred on 30 kb, whereas the HPRC event examined in detail is 62 kb, and the chromosome 20 scan found deletions up to 1.04 Mb. Recovery is declining steeply at the top of the simulated range, so extrapolating to events an order of magnitude larger is not justified by these data. The genealogies assume a single population of constant size with a uniform recombination rate, which is adequate for generating realistic linkage among carriers but does not reproduce human demographic history; carrier sets are monophyletic by construction, whereas recurrent structural variants in real data need not be.

The segmental duplication coefficient is identified between windows and its upper range is sparsely sampled, with 40 variants per size class in the high segdup stratum against 392 in the low stratum on chromosome 20. The permutation test and the leave out analyses support the direction and significance of the association but its magnitude should be treated as less precisely estimated than the size effect.

For the real data analysis the limitations are more fundamental. There is no ground truth, so the comparison establishes discordance between two representations produced by the same pipeline and cannot establish which is correct. Where a bubble contains several large events, quantities summed over its span conflate fragmentation of one event with the presence of several. Of 41 deletion events resolved across 15 bubbles on chromosome 20, all 24 that lacked a size matching record occur in only four bubbles, and 23 of them in three; every one of those bubbles has a reduced allele number, and in two the median non carrier carries as much deleted sequence as the median carrier, leaving the dispersal measure without contrast. The informative sample is therefore considerably smaller than the event count suggests. The grouping of alleles into events also depends on a breakpoint clustering tolerance, and although we report sensitivity to that choice, no threshold is uniquely correct, which is another way of saying that the unit of structural variation is not well defined in real data.

Conclusions

Pangenome graphs represent structural variation more completely than a linear reference, but that completeness is not preserved by the standard projection to VCF. The allele counts of large structural variants, particularly in segmentally duplicated regions, are fragmented by bubble decomposition; post-processing repairs much of the damage and not all of it; and the effect is a property of the representation rather than of any one graph construction pipeline. Where allele counts are the quantity of interest, they are better computed from haplotype path traversals, which recovers them almost completely, is insensitive to variant size, and costs less than the conversion it replaces. As pangenome resources grow and are used more widely for population genetic inference, the interface between graph and tabular representations deserves the same scrutiny that has been applied to the graphs themselves.

Methods

Simulation of haplotype panels with known allele counts

Benchmarking variant recovery requires ground truth that no real dataset provides. We therefore simulated haplotype panels in which the position, type, length and carrier set of every variant is known exactly, while retaining authentic genomic sequence context.

Rather than simulating sequence, we simulated the evolutionary process on top of real reference sequence. Simulated sequence lacks the repetitive architecture, segmental duplication content and transposable element composition that determine alignment difficulty, and benchmarking graph construction on such sequence would systematically understate error. Each replicate therefore begins from a 1 Mb window of GRCh38.

Genealogies were simulated with msprime 1.4.2 (Kelleher et al. 2016; Baumdicker et al. 2022) using sim_ancestry with 20 haploid samples, sequence length matched to the backbone window, recombination rate 1×10^{-8} per base per generation and population size 20,000. Because msprime scales coalescent time by the product of ploidy and population size, haploid sampling yields an expected nucleotide diversity of $2N\mu$ rather than $4N\mu$; the population size was doubled accordingly so that realised diversity matched the human genome wide average. Mutations were added with sim_mutations at 1.25×10^{-8} per base per generation. Realised diversity across replicates was $\pi = 4.9 \times 10^{-4}$ and Tajima's D = -0.055, consistent with neutral evolution at constant population size. Tree sequences were handled with tskit (Baumdicker et al. 2022).

Simulated single nucleotide variants were mapped onto the backbone by treating the backbone base as ancestral and drawing a derived base uniformly from the three alternatives. Sites falling within assembly gaps were discarded.

Structural variant placement

Structural variants were drawn from gnomAD SV v4.1 (Collins et al. 2020; Brand et al. 2024), restricted to records passing all filters and annotated as deletions, inversions or duplications, and converted to window relative coordinates. Using an empirical callset preserves realistic breakpoint placement relative to repeat structure, which randomly positioned intervals would not.

Each structural variant was assigned to a clade of the marginal genealogy at its start coordinate, so that carriers form a monophyletic group and the resulting haplotypes carry realistic linkage structure. Two sampling modes were used.

Realistic mode. The panel allele count of each variant was drawn from Binomial(20, p), where p is the variant's population allele frequency in gnomAD. Variants realising zero carriers were discarded. This reproduces the composition of a real 20 haplotype panel, in which most structural variants are small and rare.

Stratified mode. Variants were sampled across log spaced size classes centred on 100 bp, 300 bp, 1 kb, 3 kb, 10 kb and 30 kb, with four variants per class, and panel allele counts drawn uniformly from 2 to 10 independently of size. Stratification is necessary because conditioning on presence in a 20 haplotype panel selects almost exclusively for small variants: large structural variants are deleterious and therefore rare (Collins et al. 2020), so a frequency weighted sample contains too few large variants to characterise recovery as a function of size. Where a size class could not be filled from the empirical callset, intervals of the target size were placed at random gap free positions and flagged as synthetic. Across the stratified sweeps 0.4 % of variants were synthetic.

Realised lengths within each class span a range rather than taking the nominal value. Across the 2,592 stratified chromosome 20 variants the classes span 50 to 199 bp, 150 to 598 bp, 502 to 1,996 bp, 1,501 to 6,000 bp, 5,000 to 19,906 bp and 15,000 to 48,942 bp respectively, with median lengths of 85, 310,

942, 3,426, 8,000 and 22,940 bp. No variant falls below the 50 bp minimum segment length used by the graph route, so that threshold excludes nothing from the benchmark. The set comprises 1,886 deletions, 685 duplications and 21 inversions.

Because gnomAD SV is derived from short read sequencing, inversions are severely undercalled (Collins et al. 2020; Ebert et al. 2021) and a callset driven sample of a 1 Mb window typically contains none at appreciable frequency. For the inversion analysis specifically, a minimum of two inversions per size class was therefore synthesised, placed at random gap free positions so that flanking sequence context remains authentic. We note that randomly placed inversions lack the flanking inverted repeat architecture that mediates many real human inversions through non allelic homologous recombination, and that recovery measured on them should therefore be read as an upper bound.

Variants were constrained to be non overlapping with a 200 bp separation. This simplifies coordinate bookkeeping and guarantees that each event corresponds to a single graph bubble; it also means the benchmark does not test overlapping or nested structural variation, which we return to in the Discussion.

Haplotype construction and truth accounting

For each haplotype, edits were applied to the backbone in descending coordinate order so that upstream coordinates are never displaced by earlier edits. Deletions remove the interval, inversions replace it with its reverse complement, and duplications insert a tandem copy immediately following the interval.

Single nucleotide variants whose carrier set is contained within a deletion carried by the same haplotype are physically absent from every haplotype that acquired them, leaving the site monomorphic in the panel. Such sites were removed from the truth tables after haplotype construction. Retaining them would charge the graph with failing to call variants that are not present in the input sequences; across the stratified sweeps this affected approximately 3 % of simulated sites.

Haplotypes were written in PanSN format together with the unmodified backbone, labelled anc, which serves as the reference path for deconstruction and places graph derived coordinates in the same frame as the truth tables.

Window selection

Windows were selected to vary segmental duplication content while holding other properties approximately constant. Segmental duplication and RepeatMasker annotations were obtained from the UCSC Genome Browser (Kent et al. 2002), tracks genomicSuperDups (Bailey et al. 2002) and rmsk (Smit et al. 2013–2015), merged, and intersected with 1 Mb windows using BEDTools (Quinlan and Hall 2010). Windows were retained if they contained no assembly gaps (N fraction below 0.001) and had repeat content between 0.40 and 0.60, a band that encompasses most of the genome and holds general repetitive content approximately fixed so that segmental duplication varies as the principal axis. GC content was computed for the same windows with BEDTools nuc and used only as a covariate in the

confounder check, not as a selection criterion; across the retained chromosome 20 windows it ranges from 0.370 to 0.492 with a mean of 0.435.

This yielded 54 eligible windows on chromosome 20, spanning segmental duplication content from 0 to 0.253, and 61 on chromosome 16, spanning 0 to 0.974. Two independent genealogies were simulated per window on chromosome 20, giving 108 replicates and 2,592 structural variants; one genealogy per window was simulated on chromosome 16, giving 61 replicates and 1,464 structural variants. Window rather than genealogy is the unit of replication, since variants within a replicate share a graph build. Chromosome 20 was used for all method development and parameter selection. Chromosome 16 was analysed only after all parameters were fixed and served as a held out validation set; no parameter was adjusted on the basis of chromosome 16 results.

Graph construction

Graphs were built with PGGB 0.7.4 (Garrison et al. 2024), which computes all to all alignments with wfmask, induces a variation graph with seqwish (Garrison and Guarracino 2023), and normalises it with smoothxg and ODGI (Guarracino et al. 2022). Parameters were `-n 21 -t 4 -p 95 -s 5000`, with variant deconstruction against the ancestral path. Minigraph-Cactus (Hickey et al. 2024) was run through the official Docker image using `cactus-pangenome` with `--reference anc --vcf --maxCores 4`; its structural backbone is built with minigraph (Li et al. 2020).

For the builder comparison, panels were regenerated from identical seeds at all 54 chromosome 20 windows, giving 1,296 structural variants per builder on the seed 1 panels. The simulator is deterministic, and panel identity was verified by byte level comparison of the truth tables before each Minigraph-Cactus build; all 54 pairs were identical, so both builders received the same input and any difference in recovery is attributable to the builder alone.

Two routes to structural variant allele counts

VCF route. Graphs were deconstructed to VCF with `vg deconstruct` (Garrison et al. 2018), processing all snarls (Paten et al. 2018) including nested ones, then post-processed with `vcfbub` (Garrison 2022) (`-l 0 -a 100000`) and `vcfwave` (Garrison et al. 2022) (`-l 1000`). These parameters follow the Human Pangenome Reference Consortium's published pipeline (Liao et al. 2023; Lucas et al. 2026). Minigraph-Cactus applies `vcfbub` internally with the same parameters, so only `vcfwave` was applied additionally to its output. Unless stated otherwise, recovery figures are from the fully post-processed callset.

Graph route. Allele counts were computed directly from graph topology by comparing haplotype path traversals. A segment traversed exactly once by every haplotype defines an anchor, a position at which all haplotypes agree. Consecutive anchors along the reference path bracket a single top level bubble, giving a path based equivalent of snarl decomposition (Paten et al. 2018) that requires no external computation. Large deletions remove interior anchors for their carriers, so those segments cease to be

anchors and the flanking anchors bracket the whole event; this is what causes large variants to appear as single sites rather than fragmenting.

Within a bubble, each haplotype's allele is the ordered sequence of oriented segments traversed between the flanking anchors. Two haplotypes carry the same allele if and only if their traversals are identical. Orientation is retained so that an inversion registers as a distinct allele rather than matching the reference.

Traversal comparison is restricted to segments of at least 50 bp. Bubbles bracketed by distant anchors enclose many linked single nucleotide bubbles, each contributing a segment of one or a few bases; comparing raw traversals therefore makes every haplotype unique and collapses the allele count to the panel size. Recovery was insensitive to this threshold between 20 and 50 bp, both overall and within the smallest size class, and degraded above roughly 100 bp as genuine small variants began to be discarded, with the 100 bp class affected first. The threshold also sets a floor on detectable variant size: structural variants shorter than the threshold are not resolved by this route.

A bubble is reported as polymorphic when more than one distinct filtered traversal is present. Event size is defined as the larger of the length difference between traversals and the amount of sequence differing in content, since a balanced inversion has zero length difference and a length based criterion would discard every inversion. Adjacent sites sharing an identical carrier set within 50 kb are merged, since the two breakpoints of a single large event are by definition carried by the same haplotypes; carrier set identity rather than proximity is used as the merging criterion, and merging is additionally refused when any intervening site carries an allele held by the shared carrier set, since the carriers of a single deletion retain no sequence between its breakpoints.

Audit of the merge criterion

Carrier set identity is necessary but not sufficient. With 20 haplotypes the genealogy contains few distinct clades, so two independent variants can arise on the same branch and share a carrier set exactly. We audited this directly against ground truth by counting, for every merged call, how many distinct simulated variants fall within its span, and conversely how many calls overlap each simulated variant.

On chromosome 20, of 2,512 calls across the 108 replicates, 2,128 (84.7 %) contain exactly one true variant, 52 (2.07 %) span two or more and 332 (13.2 %) contain none. Of 2,592 true variants, 2,570 (99.2 %) are represented by exactly one call, 2 by more than one (residual split rate 0.0008) and 20 by none. The equivalent false fusion rate on chromosome 16 is 0.017, with 0.0082 of simulated variants left unrepresented. The criterion had genuine opportunity to err: 79 pairs of simulated variants share an identical carrier set on chromosome 20 and 63 of those lie within the 50 kb merge window, against 26 such pairs on chromosome 16. Allele count recovery was identical with and without the intervening site constraint. Merging therefore affects the reported span of a small minority of calls but not the allele

counts, which are the quantity of interest here; breakpoint coordinates from merged calls should be treated as approximate.

Staged evaluation of the conversion chain

vcfbub and vcfwave exist specifically to resolve nested bubble structure, so their contribution was measured rather than assumed. For all 108 chromosome 20 replicates the stored graph was re-deconstructed with vg deconstruct against the ancestral path to regenerate the raw VCF, and vcfbub was applied to produce the intermediate callset. The regenerated chain reproduces the original output exactly: the record count after vcfwave matches the callset produced during the original sweep for every replicate tested, which confirms that the archived parameters are the ones used.

Each of the three callsets, raw, post vcfbub and post vcfwave, was then evaluated against the same truth tables with the same matching procedure and the same recovery criteria, so that the three stages differ only in the callset supplied.

Evaluation criteria and statistical analysis

For each simulated variant, graph derived and VCF derived calls were matched by interval overlap with a tolerance of $\max(200 \text{ bp}, \text{half the variant length})$, rather than by position, because post-processing shifts coordinates and large variants may be represented at positions displaced from the true breakpoint.

Two nested criteria were applied. Under the first, a variant was scored as recovered when its true allele count appears among the alternate allele counts at any overlapping site. This criterion is deliberately permissive toward the VCF route: a site carrying many alternate alleles is scored as successful if the correct count is among them, even though selecting it would require prior knowledge of the answer. Under the second, termed readable recovery, the matched site must additionally carry at most two alleles, so that the correct count can be identified without knowing it in advance. The two criteria coincide for biallelic sites and diverge only where the decomposition has produced multiallelic records. Both were applied identically to both routes and to all three conversion stages.

Recovery was modelled as a binomial outcome with centred $\log_{10}$ variant size and segmental duplication content (scaled per 10 percentage points) as fixed effects. Standard errors were computed with a cluster robust sandwich estimator grouping by window, because variants within a window share a genomic context and a graph build and are not independent. This cluster robust generalised linear model is the primary model throughout. A replicate level linear mixed model with window as a random intercept was also fitted as a coarser check that makes the clustering explicit; it did not converge on these data, reporting a between window variance estimate on the boundary of the parameter space, and its coefficients are therefore not reported.

Collinearity between the two covariates was assessed by variance inflation factor and by the correlation between mean variant size and segmental duplication content across windows. GC content was added to the model as a potential confounder and its effect on the segmental duplication coefficient recorded.

The segmental duplication coefficient is identified between windows, and the upper part of its range is carried by few windows, so three additional robustness analyses were performed: leave one window out refitting, serial removal of the highest segmental duplication windows, and a window level permutation test (2,000 permutations) in which segmental duplication labels were shuffled across windows while preserving all within window structure. The permutation test makes no distributional assumption and is reported as the primary evidence for this coefficient.

Runtime and peak memory for the two routes were measured on a single 1 Mb window with 20 haplotypes, comparing graphsv against vg deconstruct followed by vcfbub and vcfwave on the same stored graph, with the wall clock and maximum resident set size recorded by GNU time across repeated runs.

Analyses used statsmodels 0.14.6 (Seabold and Perktold 2010), pandas 3.0.5, SciPy 1.17.1 (Virtanen et al. 2020) and NumPy 2.4.6 (Harris et al. 2020); figures were produced with Matplotlib 3.10.9 (Hunter 2007). Sequence and variant files were handled with SAMtools and BCFtools 1.24, built against HTSlib 1.23.1 (Danecek et al. 2021).

Analysis of Human Pangenome Reference Consortium data

Real data analyses used the HPRC Release 2 Minigraph-Cactus GRCh38 based pangenome, version 2.1 (Lucas et al. 2026), which extends the first draft human pangenome reference (Liao et al. 2023). Two callsets distributed with that release were compared: the raw deconstruction output, which retains the AT traversal field, and the published wave callset, from which AT has been removed by the release pipeline. Both were queried remotely over HTTPS through their tabix indices with BCFtools (Danecek et al. 2021); no file was downloaded in full, and all analyses operate on the released files as distributed.

The reference partition in this analysis comes from vg’s own bubble decomposition, recorded in the AT traversal field of the released raw callset. graphsv is not applied to HPRC data at any point. The comparison is therefore between two representations produced by the same published pipeline, with no method of ours inside it; graphsv is evaluated only against simulated panels with known truth.

Locus selection

Chromosome 20 was scanned in contiguous 5 Mb blocks of the raw callset for records carrying an alternate allele at least 20 kb shorter than the reference allele. The criterion is defined on allele length alone and is blind to allele structure, to carrier composition and to how the locus is represented in the wave callset.

Reconstruction of events from graph traversals

For each candidate bubble the AT field was parsed into one traversal per allele. Each alternate traversal was projected onto the reference traversal by retaining only those nodes that occur on the reference

path, and the largest gap in reference path index between consecutive retained nodes was taken as that allele's deleted span, with the two nodes flanking the gap as its breakpoints.

Requiring identical flanking nodes to group alleles is too strict, because haplotypes carrying one deletion often rejoin the reference a few nodes apart, which splits a single event into many near identical events and inflates any per event count. Alleles were therefore clustered by single linkage on breakpoint position, with tolerance $\max(10 \text{ reference path nodes}, 2 \% \text{ of the gap length})$. Each resulting cluster is one event; its copy number is the sum of the allele counts of its members, its size the median deleted length, and its carrier set the haplotype copies, defined as sample and phase slot pairs, whose genotype names any member allele. The number of distinct exact node pairs absorbed into each cluster and the spread of deleted lengths within it are reported alongside, and the analysis was repeated with exact node matching so that sensitivity to the clustering threshold can be assessed.

Comparison against the published callset

All wave records overlapping the bubble span were retrieved with their genotypes. Two quantities were computed for each event. The first is the set of records carrying an alternate allele whose deletion length matches the event size within $\max(200 \text{ bp}, 20 \%)$, together with the haplotype copies each such record carries; from these follow the number of matching records, the proportion of the event's carrier set recovered by the best single record and by their union, the number of records needed to cover 90 % of carriers, and the number of haplotype copies claimed by more than one record. The second is size independent: the total deleted base pairs carried by each haplotype across every record in the span, from which the excess of the event's carriers over non carriers is expressed as a fraction of the event size. A value near one indicates that the deleted sequence is represented somewhere in the span; a value near zero indicates that it is not.

Both quantities are needed because size matching alone cannot distinguish an event that is absent from one that is present at a different scale, and the second measure resolves that ambiguity without reference to record boundaries. Haplotypes with no call at a record contribute nothing to the summed deleted length, so the number of uncalled carriers is reported with each event rather than folded into it. Events were classified as recovered when a single size compatible record reproduces the carrier set, fragmented when several are required, rescaled when no record matches the event size but the carrier excess in deleted sequence is at least half the event size, and absent otherwise.

Limitations of the real data comparison

There is no ground truth in real data. These analyses compare two representations of the same graph against each other and demonstrate discordance between them; they do not establish which is correct, and they are not a test of graphsv. The simulation results address both of those questions. A bubble may legitimately contain more than one large structural variant, so counts summed over a span conflate fragmentation of a single event with the presence of several distinct events. Bubbles with markedly

reduced allele number, or in which carriers and non carriers do not differ in deleted sequence, are degenerate for this comparison and are reported separately rather than pooled.

Declarations

Availability of data and materials

All code generated for this study is available at <https://github.com/iammanjunath59-bit/graphsv> and archived at <https://doi.org/10.5281/zenodo.23083851> under the MIT licence. The repository contains graphsv, the simulation and evaluation pipeline, the statistical analysis and audits, the figure scripts, and the per variant tables from which every reported value is computed, together with a script that re-derives those values from them. Simulated haplotype panels, pangenome graphs and intermediate callsets are not deposited in the repository because they regenerate exactly from the deposited seeds; they are archived at the same Zenodo record. The human pangenome data analysed in this study are the Human Pangenome Reference Consortium Release 2 Minigraph-Cactus GRCh38 callsets, version 2.1, which are publicly available from the consortium and were read directly over HTTPS without local copies. Structural variant intervals were drawn from gnomAD SV v4.1. Reference sequence and the segmental duplication and repeat annotations were obtained from the UCSC Genome Browser.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

All code generated for this study is available at <https://github.com/iammanjunath59-bit/graphsv> and archived at <https://doi.org/10.5281/zenodo.23083851> under the MIT licence. The repository contains graphsv, the simulation and evaluation pipeline, the statistical analysis and audits, the figure scripts, and the per variant tables from which every reported value is computed, together with a script that re-derives those values from them. Simulated haplotype panels, pangenome graphs and intermediate callsets are not deposited in the repository because they regenerate exactly from the deposited seeds; they are archived at the same Zenodo record. The human pangenome data analysed in this study are the Human Pangenome Reference Consortium Release 2 Minigraph-Cactus GRCh38 callsets, version 2.1, which are publicly available from the consortium and were read directly over HTTPS without local copies. Structural variant intervals were drawn from gnomAD SV v4.1. Reference sequence and the segmental duplication and repeat annotations were obtained from the UCSC Genome Browser.

Competing interests

The author declares no competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial or not for profit sectors.

Author contributions

Manjunath M: Conceptualization, Formal analysis, Methodology, Investigation, Writing - original draft, Software, Visualization, Project administration, Resources, Writing - review & editing.

Raviraj V S: Formal analysis, Methodology, Writing - review & editing.

Shakunthala V: Writing - review & editing

Acknowledgments

The authors declare that no external funding was received for this research. We thank the Department of Studies in Genetics and Genomics, Department of studies in Zoology, University of Mysore and Institute for silkworm seed technology research, Central Silk Board, Bengaluru for providing the infrastructure.

References

1. Audano PA, Beck CR. Small polymorphisms are a source of ancestral bias in structural variant breakpoint placement. *Genome Res.* 2024;34(1):7–19. doi:10.1101/gr.278203.123
2. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, et al. Recent segmental duplications in the human genome. *Science.* 2002;297:1003–1007. doi:10.1126/science.1072047
3. Baumdicker F, Bisschop G, Goldstein D, Gower G, Ragsdale AP, Tsambos G, et al. Efficient ancestry and mutation simulation with msprime 1.0. *Genetics.* 2022;220(3):iyab229. doi:10.1093/genetics/iyab229
4. Brand H, Zhao X, Fu J, Collins R, Wong I, Liao C, et al. O11: an atlas of 1.2M structural variants across global populations in the Genome Aggregation Database (gnomAD). *Genet Med Open.* 2024;2(Suppl 1):101015. doi:10.1016/j.gimo.2024.101015 [meeting abstract]
5. Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Francioli LC, et al. A structural variation reference for medical and population genetics. *Nature.* 2020;581:444–451. doi:10.1038/s41586-020-2287-8
6. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. *GigaScience.* 2021;10(2):giab008. doi:10.1093/gigascience/giab008
7. Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science.* 2021;372:eabf7117. doi:10.1126/science.abf7117

8. Ebler J, Ebert P, Clarke WE, Rausch T, Audano PA, Houwaart T, et al. Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nat Genet.* 2022;54(4):518–525. doi:10.1038/s41588-022-01043-w
9. Garrison E. vcfbub: popping bubbles in vg deconstruct VCFs. Software. 2022. <https://github.com/pangenome/vcfbub>
10. Garrison E, Guarracino A. Unbiased pangenome graphs. *Bioinformatics.* 2023;39(1):btac743. doi:10.1093/bioinformatics/btac743
11. Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat Biotechnol.* 2018;36(9):875–879. doi:10.1038/nbt.4227
12. Garrison E, Kronenberg ZN, Dawson ET, Pedersen BS, Prins P. A spectrum of free software tools for processing the VCF variant call format: vcflib, bio-vcf, cyvcf2, hts-nim and slivar. *PLoS Comput Biol.* 2022;18(5):e1009123. doi:10.1371/journal.pcbi.1009123
13. Garrison E, Guarracino A, Heumos S, Villani F, Bao Z, Tattini L, et al. Building pangenome graphs. *Nat Methods.* 2024;21:2008–2012. doi:10.1038/s41592-024-02430-3
14. Guarracino A, Heumos S, Nahnsen S, Prins P, Garrison E. ODGI: understanding pangenome graphs. *Bioinformatics.* 2022;38(13):3319–3326. doi:10.1093/bioinformatics/btac308
15. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. *Nature.* 2020;585:357–362. doi:10.1038/s41586-020-2649-2
16. Hickey G, Heller D, Monlong J, Sibbesen JA, Sirén J, Eizenga J, et al. Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biol.* 2020;21:35. doi:10.1186/s13059-020-1941-7
17. Hickey G, Monlong J, Ebler J, Novak AM, Eizenga JM, Gao Y, et al. Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nat Biotechnol.* 2024;42:663–673. doi:10.1038/s41587-023-01793-w
18. Hunter JD. Matplotlib: a 2D graphics environment. *Comput Sci Eng.* 2007;9(3):90–95. doi:10.1109/MCSE.2007.55
19. Kelleher J, Etheridge AM, McVean G. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Comput Biol.* 2016;12(5):e1004842. doi:10.1371/journal.pcbi.1004842
20. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. *Genome Res.* 2002;12:996–1006. doi:10.1101/gr.229102
21. Li H, Feng X, Chu C. The design and construction of reference pangenome graphs with minigraph. *Genome Biol.* 2020;21:265. doi:10.1186/s13059-020-02168-z
22. Liao WW, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, et al. A draft human pangenome reference. *Nature.* 2023;617:312–324. doi:10.1038/s41586-023-05896-x
23. Lucas JK, Hebbbar P, Liao WW, Macias-Velasco JF, Novak AM, Asri M, et al. HPRC2: a human pangenome reference with near-complete coverage of common genetic variation. *bioRxiv.* 2026.

doi:10.64898/2026.07.21.739710

24. Paten B, Eizenga JM, Rosen YM, Novak AM, Garrison E, Hickey G. Superbubbles, ultrabubbles, and cacti. *J Comput Biol.* 2018;25(7):649–663. doi:10.1089/cmb.2017.0251
25. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.* 2010;26(6):841–842. doi:10.1093/bioinformatics/btq033
26. Seabold S, Perktold J. statsmodels: econometric and statistical modeling with Python. In: *Proceedings of the 9th Python in Science Conference*; 2010. p. 92–96. doi:10.25080/Majora-92bf1922-011
27. Sirén J, Monlong J, Chang X, Novak AM, Eizenga JM, Markello C, et al. Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science.* 2021;374(6574):abg8871. doi:10.1126/science.abg8871
28. Smit AFA, Hubley R, Green P. RepeatMasker Open-4.0. 2013–2015. <http://www.repeatmasker.org>
29. Tan A, Abecasis GR, Kang HM. Unified representation of genetic variants. *Bioinformatics.* 2015;31(13):2202–2204. doi:10.1093/bioinformatics/btv112
30. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods.* 2020;17:261–272. doi:10.1038/s41592-019-0686-2
31. Vollger MR, Dishuck PC, Harvey WT, DeWitt WS, Guitart X, Goldberg ME, et al. Increased mutation and gene conversion within human segmental duplications. *Nature.* 2023;617:325–334. doi:10.1038/s41586-023-05895-y
32. Zhou Y, Zhang Z, Bao Z, Li H, Lyu Y, Zan Y, et al. Graph pangenome captures missing heritability and empowers tomato breeding. *Nature.* 2022;606:527–534. doi:10.1038/s41586-022-04808-9

Figures

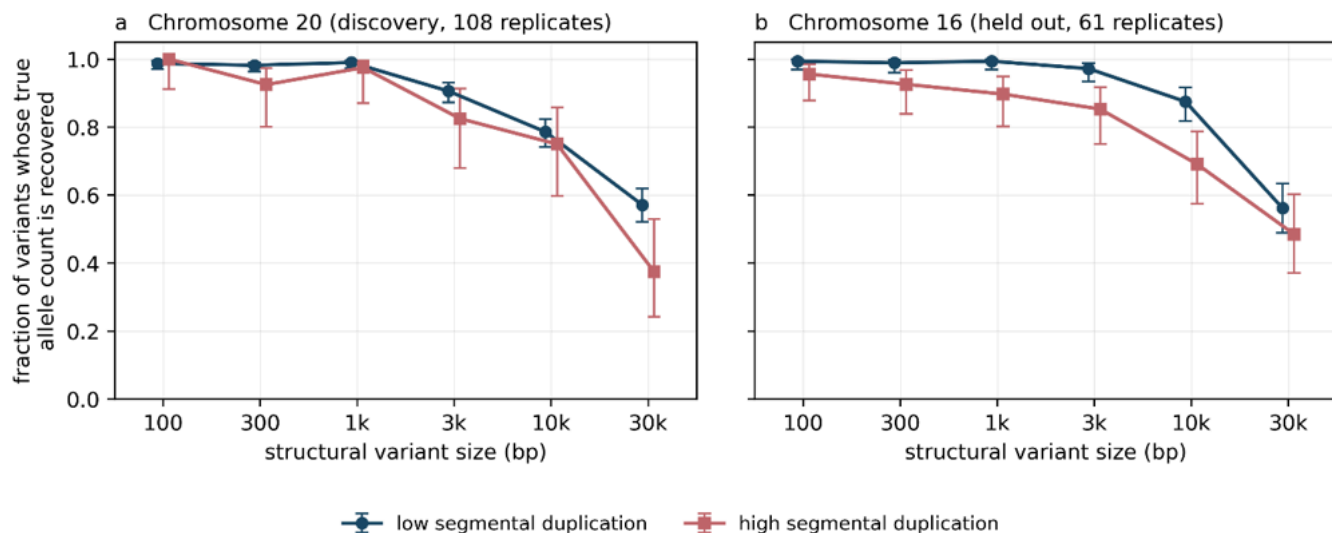


Figure 1

Allele count recovery declines with structural variant size and with segmental duplication content.

Recovery of the true allele count of simulated structural variants from the VCF route, in graphs built with PGGB, as a function of variant size. Variants are split into windows of low and high segmental duplication content. **a** Chromosome 20, the discovery set: 108 replicates across 54 windows, 2,592 variants, 392 and 40 variants per size class in the low and high strata respectively. **b** Chromosome 16, analysed only after all parameters were fixed: 61 replicates, 1,464 variants, 176 and 68 per class. A variant counts as recovered when its true allele count appears among the alternate allele counts at an overlapping site. Points are the observed fraction and bars are Wilson binomial intervals; these do not account for clustering of variants within windows, and the interval estimates reported in the text use a cluster robust estimator. Recovery falls from 0.988 at 100 bp to 0.553 at 30 kb on chromosome 20, and the two covariates compound at the largest sizes, where recovery is 0.571 in low segdup windows against 0.375 in high segdup windows.

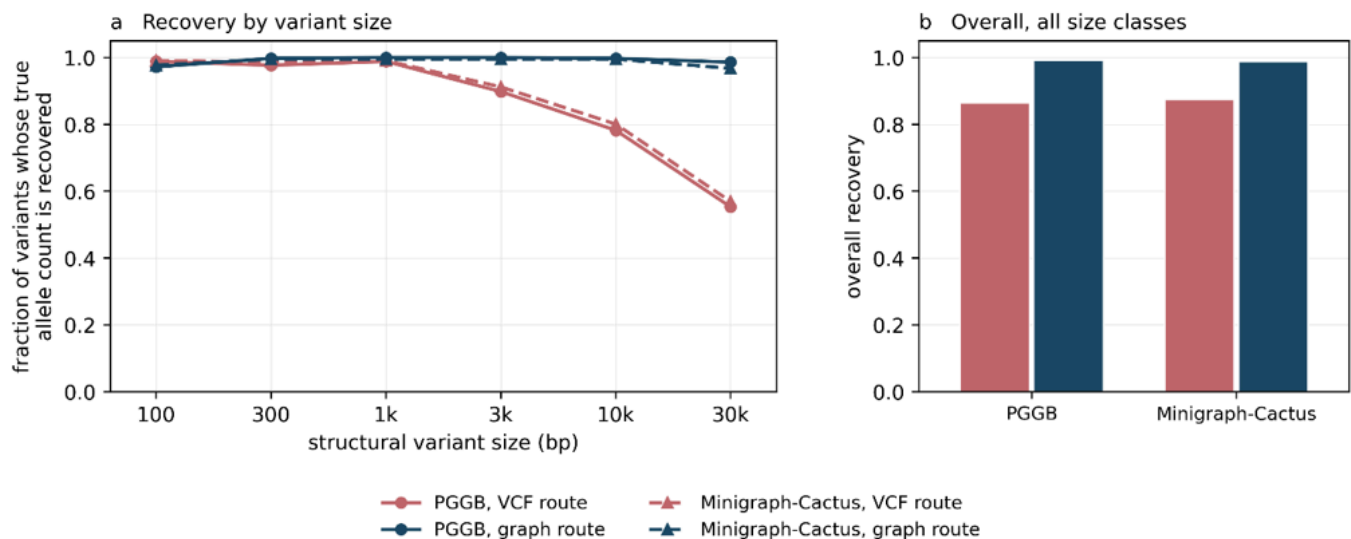


Figure 2

The loss is a property of the VCF representation rather than of the graph builder. a Allele count recovery against variant size for the VCF route and for allele counts computed directly from haplotype path traversals, in graphs built by PGGB and by Minigraph-Cactus from byte identical haplotype panels. **b** The same quantities aggregated across all size classes. PGGB values are from 108 replicates and 2,592 variants, Minigraph-Cactus values from the 54 seed 1 panels and 1,296 variants; all 54 truth tables were verified byte identical between the two arms before building. The two routes agree for small variants and diverge as size increases, reaching a difference of 0.433 in the 30 kb class for PGGB. Overall recovery is 0.865 for the VCF route and 0.992 for the graph route in PGGB graphs, and 0.875 and 0.988 respectively in Minigraph-Cactus graphs.

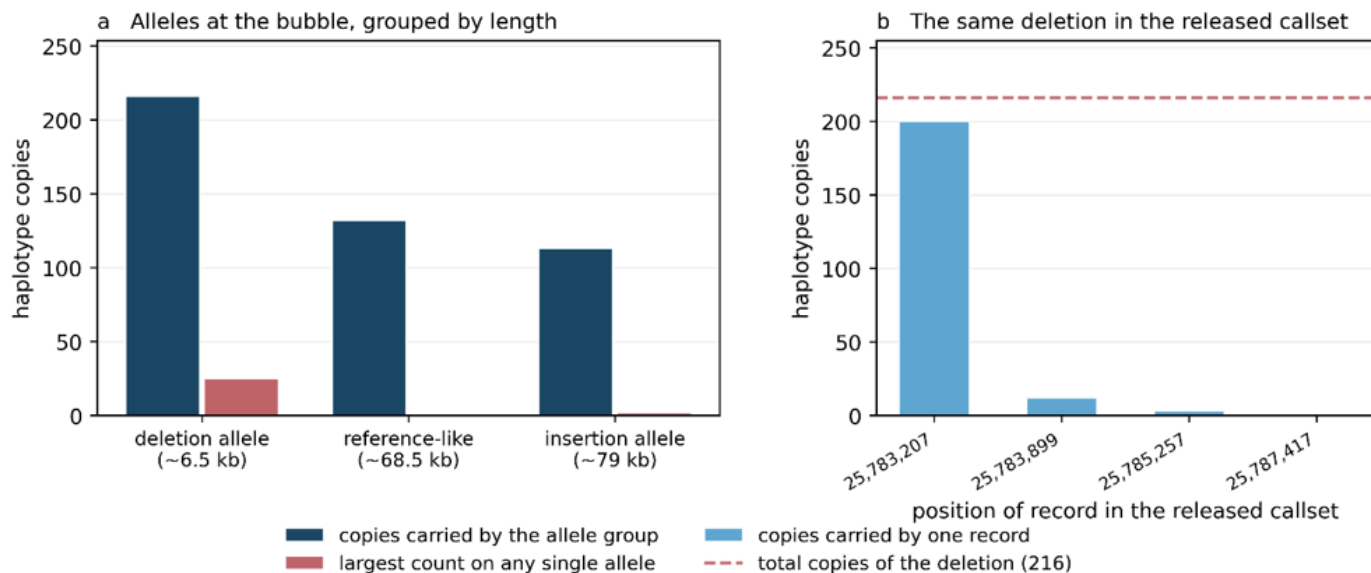


Figure 3

A common deletion in the HPRC Release 2 pangenome is split across alleles in the raw decomposition and across records in the released callset. The top level bubble at chr20:25,783,205 spans 68,466 bp of GRCh38 and carries 328 alternate alleles across 461 haplotypes. **a** Alleles grouped by alternate length into three classes, with the haplotype copies carried by each group beside the largest count attached to any single allele within it. The 84 alleles of the deletion class share one pair of breakpoint nodes and together carry 216 copies, a frequency of 0.469, while the largest single allele count is 25, a frequency of 0.054. **b** The same deletion in the released wave callset, where four records carry 200, 12, 3 and 1 haplotype copies. These sum to exactly 216 and no haplotype copy appears in more than one record, so the carriers are partitioned rather than duplicated or lost. The partition used as the reference here is vg's own, recorded in the AT traversal field of the raw callset; graphsv is not applied to these data.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementaryfile1.docx](#)
- [Softwareversions.docx](#)