Small allelic variants are a source of ancestral bias in structural variant breakpoint placement

Running Title: Ancestral bias alters structural variant breakpoints

Peter A. Audano[1] and Christine R. Beck[1,2]

1. The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA
2. Department of Genetics and Genome Sciences, Institute for Systems Genomics, University of Connecticut Health Center, Farmington, CT, USA

Corresponding author: christine.beck@jax.org

Abstract

High-quality genome assemblies and sophisticated algorithms have increased sensitivity for a wide range of variant types, and breakpoint accuracy for structural variants (SVs, ≥ 50 bp) has improved to near basepair precision. Despite these advances, many SVs in unique regions of the genome are subject to systematic bias that affects breakpoint location. This ambiguity leads to less accurate variant comparisons across samples, and it obscures true breakpoint features needed for mechanistic inferences. To understand why SVs are not consistently placed, we re-analyzed 64 phased haplotypes constructed from long-read assemblies released by the Human Genome Structural Variation Consortium (HGSVC). We identified variable breakpoints for 882 SV insertions and 180 SV deletions not anchored in tandem repeats (TRs) or segmental duplications (SDs). While this is unexpectedly high for genome assemblies in unique loci, we find read-based callsets from the same sequencing data yielded 1,566 insertions and 986 deletions with inconsistent breakpoints also not anchored in TRs or SDs. When we investigated causes for breakpoint inaccuracy, we found sequence and assembly errors had minimal impact, but we observed a strong effect of ancestry. We confirmed that polymorphic mismatches and small indels are enriched at shifted breakpoints and that these polymorphisms are generally lost when breakpoints shift. Long tracts of homology, such as SVs mediated by transposable elements, increase the likelihood of imprecise SV calls and the distance they are shifted. Tandem Duplication (TD) breakpoints are the most heavily affected SV class with 14% of TDs placed at different locations across haplotypes. While graph genome methods normalize SV calls across many samples, the resulting breakpoints are sometimes incorrect, highlighting a need to tune graph methods for breakpoint accuracy. The breakpoint inconsistencies we characterize collectively affect ~5% of the SVs called in a human genome and underscore a need for algorithm development to improve SV databases, mitigate the impact of ancestry on breakpoint placement, and increase the value of callsets for investigating mutational processes.

Introduction

The human reference genome (International Human Genome Sequencing Consortium, 2001; Schneider et al., 2017) hosts annotations including genes (Frankish et al., 2021; O'Leary et al., 2016), regulatory regions (Encode Project Consortium, 2012; Encode Project Consortium et al., 2020), and repeats (Bailey et al., 2002; Benson, 1999; Smit, 2013-2015), and it has become a universal coordinate system for describing genetic alterations across populations (Abel et al., 2020; Audano et al., 2019; Beyter et al., 2021; Collins et al., 2020; Ebert et al., 2021; International HapMap et al., 2007; Karczewski et al., 2020; Sudmant et al., 2015; The 1000 Genomes Project Consortium, 2015) and diseases (ICGC TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020; Taliun et al., 2021; Turner et al., 2017). New high-quality references are emerging for humans (Nurk et al., 2022) and a growing number of other species (Alonge et al., 2020; Ferraj et al., 2023; Jebb et al., 2020; Li et al., 2023; Mao et al., 2021; Mouse Genome Sequencing Consortium et al., 2002), which play fundamental roles in modern genomics.

Variant discovery is largely based on aligning reads or assemblies to a reference genome. This is used to identify single nucleotide variants (SNVs), small insertions and deletions (indels), and structural variants (SVs) including insertions and deletions ≥ 50 bp, inversions, complex rearrangements, and chromosomal translocations. Imprecise SV breakpoints affect comparisons across samples, and while new methods are improving comparisons (Ebert *et al.*, 2021; English et al., 2022; Kirsche et al., 2021), error-free merging across many haplotypes has not yet been attained. Additionally, breakpoint features such as microhomology and nearby variants in-*cis* are important signatures for predicting mechanisms of formation (Beck et al., 2015; Carvalho and Lupski, 2016; Carvalho et al., 2011; Vogt et al., 2014). Repetitive sequences often mediate SVs and can make the determination of precise breakpoints challenging.

Recent advances in sequencing technology are now generating longer and more accurate reads capable of reaching into repetitive structures and spanning larger SVs. As a result, many new SV loci have been discovered, and SV yield per sample has increased from less than 10,000 SVs per genome to more than 25,000 (Audano *et al.*, 2019; Chaisson et al., 2015; Ebert *et al.*, 2021). Moreover, long-reads routinely reveal the full sequence of SVs, which was not previously attainable. In more recent years, long-read phased assemblies have become a critical component for producing complete and accurate variant callsets (Chaisson et al., 2019; Ebert *et al.*, 2021; Garg et al., 2021; Liao et al., 2023). These advances enable more complete transposable element (TE) analysis, improve genotyping in short-read samples, and support new biological insights (Ebert *et al.*, 2021; Ebler et al., 2022; Rozowsky et al., 2023).

Modern references are a single theoretical human haplotype creating alignment biases when reads are mapped to non-reference alleles, which can be difficult to mitigate (Brandt et al., 2015; Degner et al., 2009; Eizenga et al., 2020). To support mapping and variant calling across diverse genomes, the Human Pangenome Reference Consortium (HPRC) is developing graph-based references encompassing many haplotypes simultaneously (Liao *et al.*, 2023). While in-graph haplotypes can be directly detected, variants absent from the graph reference still rely on calling differences between the sample and a graph path. Therefore, challenges with linear reference analyses will ultimately translate to graphs, especially for rare and somatic events often associated with disease (Nattestad et al., 2018; Rausch et al., 2023; Sakamoto et al., 2020; Vogt *et al.*, 2014; Wahlster et al., 2021). However, in-graph SVs are represented as a unique "bubble" in graph space with a common breakpoint across samples, and so ambiguity with merging across independent haplotypes may be eliminated, although breakpoint accuracy has not been assessed.

While contiguous high-accuracy assemblies are becoming routine, we find that SV breakpoints are still inconsistently placed across phased haplotypes, and many breakpoints do not represent the true site of rearrangements, potentially impeding downstream analyses. To quantify the effect on modern long-read variant discovery approaches, we re-analyze a recent callset from 64 phased haplotypes recently released by the Human Genome Reference Consortium (HGSVC) (Ebert *et al.*, 2021). With pangenomes recently released by the Human Pangenome Reference Consortium (HPRC) (Liao *et al.*, 2023), we identify discordance between linear- and graph–based-reference approaches. We determine reasons why breakpoints can differ between assemblies and suggest approaches for improving both mechanistic inference and variant comparisons across samples.

Results

Breakpoint offsets are prevalent in long-read SV callsets

We examined breakpoint placement for SVs across 64 phased haplotypes derived from 32 diverse samples released by the HGSVC (Ebert *et al.*, 2021). In that study, variants were called independently on each assembled haplotype against the GRCh38 reference using minimap2 (Li, 2018) and merged to a multi-haplotype, nonredundant callset. For each pair of haplotypes (2,016 combinations of 64 haplotypes), we find an average of 20% of insertions and 15% of deletions have different breakpoints between the pair. When SVs anchored in tandem repeats (TRs) and segmental duplications (SDs) are excluded (Methods), we find 4.4% of insertions and 1.7% of deletions on average have different breakpoints (**Table S1**). In this paper, we refer to "unique loci" as regions outside TRs and SDs where large and highly repetitive structures often produce ambiguous alignments. We include transposons in our unique loci because the 64

assemblies we are investigating (average n50 > 19.5 Mbp) are capable of spanning full-length human TEs, which cluster around 300 bp and 6 kbp.

Inconsistent breakpoints in unique regions affect a small number of variants per haplotype pair, but the effect across multiple haplotypes and samples is greater. In the merged callset across all 64 haplotypes, we find 5.9% insertions and 3.1% deletions in unique loci disagree on breakpoint location (**Table 1**). While many of these differences are small, insertions vary by a median of 2.2 bp and deletions by 4.9 bp resulting in a non-trivial effect on SV representation (**Fig 1A, Table 1**).

Finally, the number of distinct breakpoints for each variant does not scale linearly with the number of haplotypes harboring the SV (AC: allele count) (**Fig 1B**). This suggests that variant breakpoints are placed consistently across many haplotypes, but are affected for a subset of haplotypes.

Table 1: Summary of differential breakpoints in the merged callset. Variants in unique regions (No TR/SD) have the fewest breakpoint offsets, which grows as SVs in more complex loci are included. TR: Tandem Repeat, SD: Segmental Duplication, N: Number of variants, Diff: Variants with different offsets in at least one haplotype, Med: Median breakpoint offset.

	Insertions				Deletions			
	N	Diff	Diff %	Med bp	N	Diff	Diff %	Med bp
No TR/SD	14,961	882	5.9%	2.2	5,804	180	3.1%	4.9
SD No TR	2,609	440	16.9%	3.5	1,922	402	20.9%	5.0
All	60,716	19,589	32.3%	19.4	38,442	10,641	27.7%	22.0

Diversity is the main driver of differential breakpoint placement

To examine whether random sequence errors may affect SV quality, we compared CLR (21 genomes) with HiFi (11 genomes) in the HGSVC callset. We find a marginally significant enrichment for differential breakpoints in CLR vs HiFi for insertions (4.40% vs 4.29%, $p = 0.025$, Student's t-test) and no enrichment for deletions (1.75% vs 1.77%, $p = 0.52$, Student's t-test),

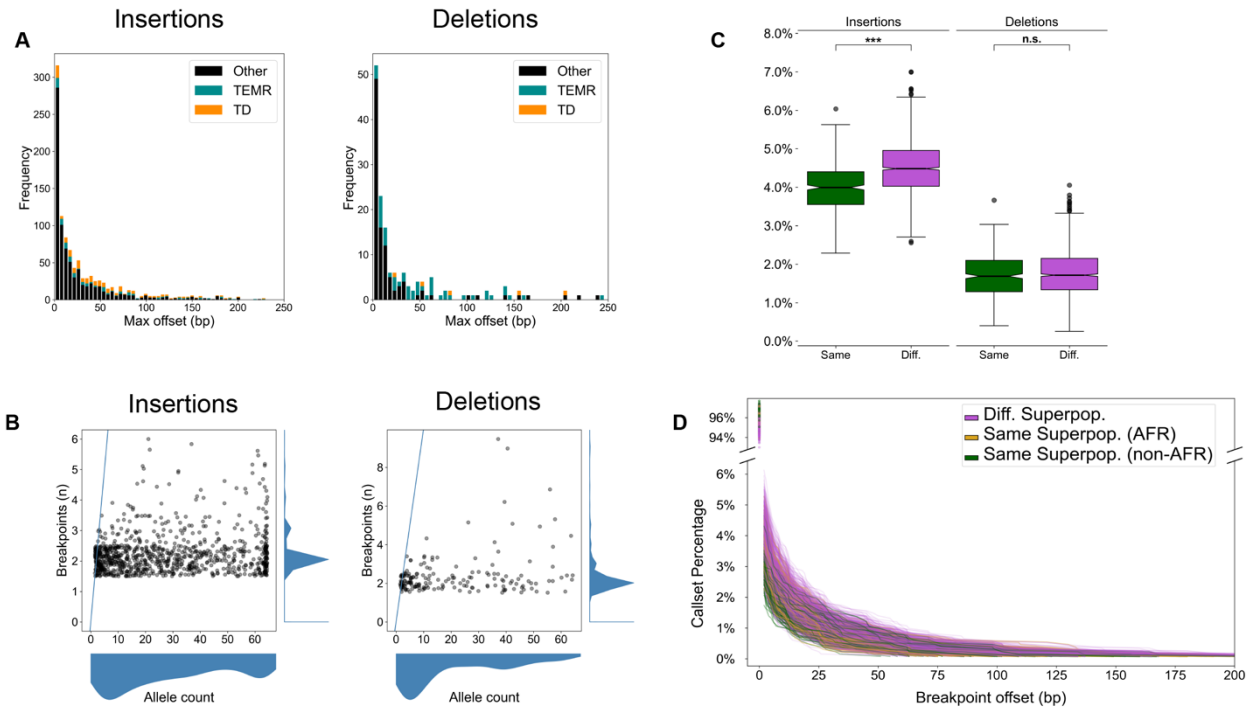

Figure 1: Breakpoint offsets reflect population structure. (A) Maximum offset distances for variants in the merged callset. For SVs identified in more than one haplotype, the maximum offset is the difference between the left- and right-most breakpoints. TEMR: Transposable element mediated rearrangements. TD: Tandem duplications. **(B)** The number of unique breakpoints for each variant (vertical axis) does not scale with the number of haplotypes (horizontal axis). A blue line represents the x = y diagonal. Scatterplot points were jittered in each axis uniformly from -0.5 to 0.5 to show density. **(C)** For any pair of haplotypes, the proportion of offset SVs is stratified by same superpopulation (green) or different superpopulation (violet). The difference in means is significant for both insertions and deletions (Student's T-test of means), but a greater effect is seen for insertions. Notches indicate a 95% confidence interval around the median. n.s: Not significant, *: 1e-3 < p ≤ 1e-2, **: 1e-4 < p ≤ 1e-3, ***: p ≤ 1e-4. **(D)** A cumulative distribution of breakpoint offset for all haplotype pairs. Most variants in both haplotypes share the same breakpoint (upper y-axis). For variants with at least 1 bp offset (lower y-axis), the cumulative proportion of matched calls decreases with increasing breakpoint distance. When both samples come from a different superpopulation (violet), larger differences between breakpoints are observed than when haplotypes come from the same superpopulation (green). When both haplotypes come from African samples (gold), breakpoint distances are elevated, but to a lesser extent than different ancestral backgrounds.

which we confirmed with permutation tests (p = 0.012 insertions, p = 0.74 deletions, 100,000 permutations).

We next asked whether sequence variation affected placement. To investigate this, we stratified the callset by ancestry across 64 HGSVC haplotypes derived from all five 1000 Genomes

ancestral superpopulations composed of African, admixed American, East Asian, European, and South Asian ancestry. We observed that variant breakpoints differed more often when a pair of samples were derived from different superpopulations for insertions (4.49% vs 3.99%, $p = 2.44\times10^{-40}$, Welch's t-test, Cohen's D = 0.73) (**Fig 1C**). Deletions were also increased, but the effect did not reach significance (1.76% vs 1.71%, $p = 0.069$, Welch's t-test) (**Fig 1C**). Furthermore, there is a noticeable increase in offset distance when haplotypes are derived from different ancestral backgrounds (**Fig 1D**); we confirmed these results with permutation tests ($p < 1\times10^{-5}$ insertions, $p = 0.041$ deletions, 10,000 permutations). Our results suggest that allelic polymorphisms in or near SVs are a driver of breakpoint differences, which reveals a source of ancestral bias in modern SV callsets.

Breakpoint offsets are more prevalent with TE-mediated SVs

Transposable elements (TEs) create tracts of homology throughout the genome resulting in TE-mediated rearrangements (TEMRs) (Balachandran et al., 2022; Han et al., 2008; Sen et al., 2006). TEs from the same family have highly similar sequences, and so there are many choices for breakpoint placement along TE copies (**Fig. 2A**). While TEs may provide the homology necessary for duplications and deletions by non-allelic homologous recombination (NAHR), most exhibit only short tracts of breakpoint homology and appear to be mediated by other repair processes (Balachandran *et al.*, 2022). Therefore, accurately placing SV breakpoints within TEMRs is essential for understanding the mutational mechanisms underlying their formation (Morales et al., 2015).

In the merged HGSVC SV callset, we find 112 SV insertions and 119 SV deletions with differential breakpoints in unique loci are likely TEMRs (8.5% and 20.4% of differential variants, respectively) (Methods). We find TEMR insertions were significantly enriched for offset breakpoints (odds ratio (OR) = 4.18, $p = 3.17\times10^{-25}$, Fisher's exact test (FET)), as were TEMR

156 deletions (OR = 3.20, $p = 1.55\times10^{-11}$, FET). TE homology is also responsible for larger distances
157 between breakpoints across haplotypes for insertions (15.17 vs 2.50 bp, $p = 1.45\times10^{-8}$, Welch's
158 T-test), but the breakpoint difference was did not reach significance for deletions (46.71 vs
159 10.93 bp, $p = 0.065$, Welch's T-test) (**Fig 2B**).

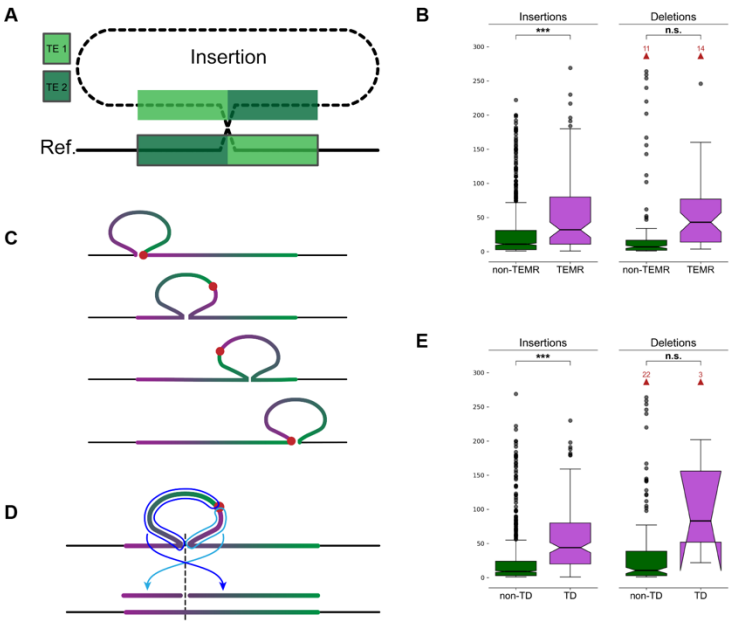

Figure 2: Breakpoints shift through homology and alter SV representation. (A) A non-reference sequence with chimeric TE copies (light green and dark green) at the breakpoints. **(B)** The maximum distance between differential breakpoints across haplotypes is greater in TEMR variants with a significant difference seen for insertions. **(C)** The sequence of a tandem duplication is rotated as the alignment shifts along the reference copy. The true breakpoint (red dot) is located at the insertion breakpoint if the duplicate copy is aligned to the left or right end of the reference copy (top and bottom examples), and it is embedded within the insertion otherwise creating a chimera of duplication copies. **(D)** Mapping a rotated TD to the reference occurs in two pieces separated at the TD breakpoint (light and dark blue arrows) where each piece maps to one side of the reference insertion site (dashed line). Alignment programs often miss one or both fragments. **(E)** Maximum breakpoint offsets in TDs and non-TD SVs. Horns extending downward on the TD deletions indicates that the 95% confidence interval for the median extends below the bottom quartile. (**B**, **E**) *p*-values are generated from T-tests of the mean. Notches indicate a 95% confidence interval around the median. Red arrows and numbers indicate the number of outlier points above the horizontal axis maximum. n.s.: Not significant, *: 1e-3 < $p \leq$ 1e-2, **: 1e-4 < $p \leq$ 1e-3, ***: $p \leq$ 1e-4.

160

Tandem duplications are heavily affected by differential breakpoints

Tandem duplications (TDs) are a common SV type where a duplicate copy is inserted adjacent to its template. TDs may be driven by existing homology, such as NAHR, or occur in regions with little to no homology (Arlt et al., 2009; Lee et al., 2007; Li et al., 2020; Menghi et al., 2016; Willis et al., 2017). With short reads, TDs are detected by elevated copy number of the duplicated sequence combined with paired-end evidence at the duplication breakpoint, revealing the duplicated reference region (Alkan et al., 2011). However, long-read callers often call TDs as insertions, especially when assemblies are used. A TD is highly homologous with itself posing a significant problem for alignment algorithms because there are many valid choices for the breakpoint placement. If the breakpoint is not placed on one end of a reference copy, the insertion sequence contains a chimera of both duplication copies and the true breakpoint is embedded somewhere inside the insertion sequence (**Fig 2C**). Annotating a TD should be as easy as re-aligning the insertion sequence to the reference and determining if it maps adjacent to the insertion breakpoint, however, rotated TDs align in two separate fragments (**Fig. 2D**), and current alignment programs often miss one or both fragments. To better annotate SVs as TDs, we re-aligned SV sequences with BLAST (Altschul et al., 1990) (Methods). We found 1,843 SV insertions were TDs, of which 261 (14.2%) were shifted and rotated leading to the duplication mapping to two separate BLAST records on each side of the insertion site. We found 17 reference TDs with a deleted copy, of which 8 (47.1%) were shifted and rotated mapping to both sides of the deletion SV.

We find that TD insertions are more likely to have differential breakpoints (OR = 0.55, p = 1.45×10^{-9}, FET), but the effect on TD deletions is small (OR = 0.05, p = 3.14×10^{-7}, FET). Because homology runs across the full length of a TD, we observe greater average offset distances for insertions (9.37 bp TD vs 2.20 bp non-TD, p = 1.07×10^{-13}, Welch's t-test, Cohen's D = 0.45). A large increase in distance for deletions failed to reach significance (741.9 bp TD vs

12.8 bp non-TD, p = 0.19, Welch's t-test, Cohen's D = 2.23) (**Fig. 2E**). These results appear to suggest that TD deletions are highly susceptible to breakpoint shifts, however, the low number of these events and the large range of offsets across all deletions make these observations difficult to validate.

Small polymorphisms surround offset SV breakpoints

Differential breakpoints occur in regions with tracts of homology, such as TEMRs and TDs. In cases of perfect homology, the actual breakpoint could have occurred anywhere in the homologous region. By convention, many aligners, such as minimap2 (Li, 2018), push breakpoints to the left yielding more consistent variant calls across haplotypes. As we have observed, variation in breakpoint placement increases when haplotypes are derived from different ancestral superpopulations, therefore, we reasoned that small allelic polymorphisms around SV breakpoints might influence alignments.

To identify small polymorphisms at SV breakpoints that might influence alignments, we extracted the offset region around breakpoints from each haplotype assembly and compared them (**Fig 3A**) (Methods). For SV insertions in unique regions with shifted breakpoints, we find on average 5.0 small variants on the left breakpoint vs 5.2 on the right breakpoint ($p = 1.62\times10^{-10}$, Welch's t-test) (**Fig 3B**).

When polymorphisms occur in homologous regions, it creates differences between the haplotype and the reference sequence, which penalizes the alignment score for the haplotype more diverged from the reference. If the variant can be shifted across a homologous region, a better alignment score may be achieved by moving the breakpoint such that small polymorphisms are pushed into the unaligned insertion sequence. As a consequence, we observe a large peak of small polymorphisms on the upstream breakpoint shifted 1 bp inside the

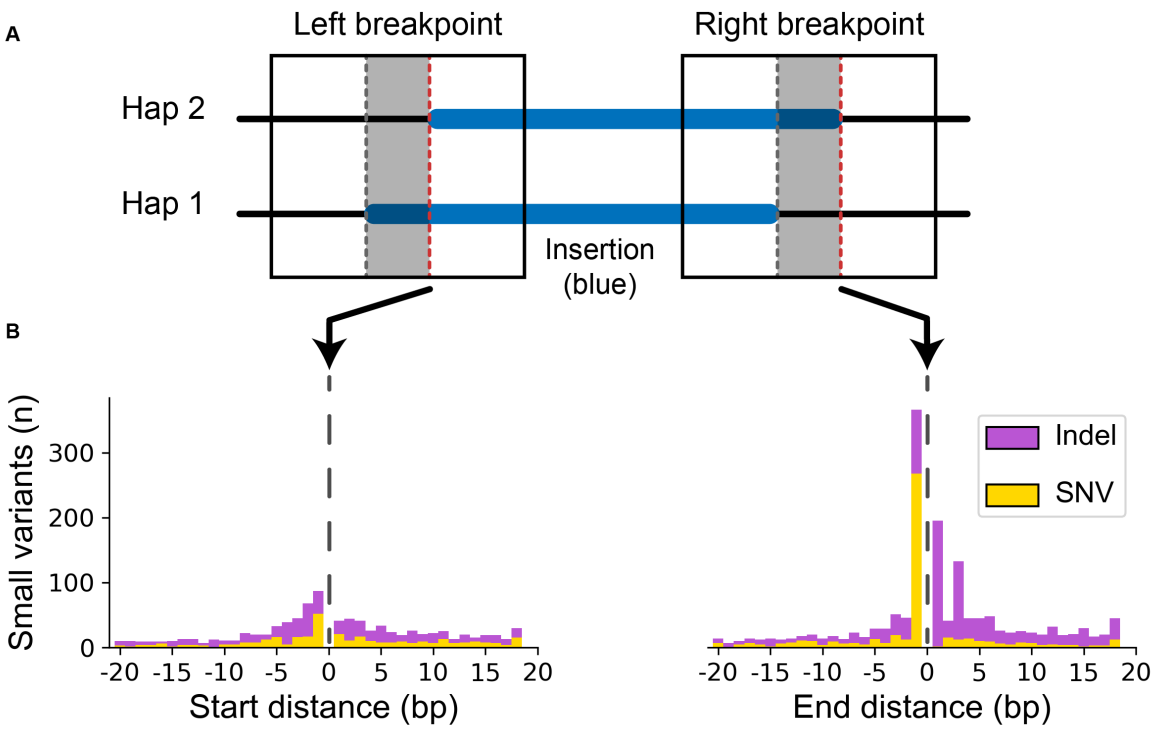

Figure 3: Small variants accumulate at differential SV breakpoints. (A) SV insertions with different breakpoints (blue) in each haplotype pair were retrieved. Sequence around the left and right breakpoints was extracted (solid box) for both haplotypes including the differential locus (gray area between dashed lines) and 50 bp upstream and downstream. The red dashed line marks the start and end position of the right-shifted variant. **(B)** Small variants accumulate at the upstream and downstream edges of the right-shifted variants where 0 is the red line in part (A).

insertion (**Fig 3B**). Not only does this drive breakpoint disagreements, but these small polymorphisms near SVs are systematically removed from variant callsets.

Breakpoint homology annotations change with breakpoint placement

SVs are often mediated by large tracts of homology during their formation through NAHR requiring more than 100 bp of perfect homology, double-strand break repair pathways mediated by tracts of microhomology from 1 to 10 bp, non-homologous end joining (NHEJ) requiring no breakpoint homology, and alternative end-joining (alt-EJ) requiring little or no microhomology (reviewed in Carvalho and Lupski (2016) and Iliakis et al. (2015)). Mobile element insertions (MEIs) create homology in the form of target-site duplications (TSDs), which is an important

220 annotation for distinguishing true MEI polymorphisms from other SVs containing MEI sequence
221 (Ebert *et al.*, 2021; Zhou et al., 2020). Accurately detecting breakpoint homology is an important
222 predictor of SV mechanism and a useful quality metric for SV callsets, however, the effect of
223 differential breakpoints on microhomology has not been investigated.

224 We used a recent PAV addition to estimate microhomology for all SV breakpoints
225 (Balachandran *et al.*, 2022) (Methods). For each SV, we find that the number of breakpoints
226 increases the number of different microhomology calls for insertions ($\rho = 0.72$, $p < 1\times10^{-100}$,
227 Spearman rank-order correlation (Spearman)) and deletions ($\rho = 0.87$, $p < 1\times10^{-100}$, Spearman)
228 indicating that breakpoint changes affect homology annotations in almost all cases (**Fig. S1**).
229 For insertions with consistent breakpoints (n = 6,855), microhomology annotations varied by
230 2.16 bp on average, which rises to 21.91 bp on average with inconsistent breakpoints (n = 725)
231 ($p = 9.46\times10^{-15}$, Welch's t-test, Cohen's D = 0.43). We see a similar effect on deletion
232 microhomology, which varies by 0.01 bp across haplotypes with consistent breakpoints (n =
233 3,399) and rises to 19.27 bp across haplotypes with inconsistent breakpoints (n = 172) ($p =$
234 1.01×10^{-16}, Welch's t-test, Cohen's D = 1.77) (**Fig. S2**).

235 As a result of imprecise breakpoints, actual breakpoint homologies necessitate manual
236 reconstruction, which is tedious task and cannot easily scale with modern whole-genome
237 analyses. Therefore, precise mechanisms are difficult to annotate at scale. For example, while
238 SVs mediated by mobile elements with at least 85% identity are generally thought to be
239 mediated by non-allelic homologous recombination (NAHR) (Lam et al., 2010), a closer
240 examination of breakpoints using modern long-read data shows that at least 20% have
241 breakpoint features inconsistent with NAHR (Balachandran *et al.*, 2022).
242

Read-based approaches have less consistent breakpoints

We examined the effects of offset breakpoints from aligned reads using PBSV. In our callset, 11,906 SV insertions and 5,501 SV deletions in unique loci were callable across the HGSVC samples. We find that 13% of insertions and 18% of deletions are offset across samples when called from read alignments, which is higher than 6% insertions and 3% deletions we observe from assembly-based callsets for the same SVs. We hypothesize that assemblies are more consistent because a single polished representation of the region is aligned where individual reads may be subject to more systematic bias, for example read errors and SVs on the edges of individual reads.

While short-read callers typically rely on read alignments, some produce breakpoint assemblies and may improve breakpoint accuracy. A recent study of TEMRs (Balachandran *et al.*, 2022) finds that MANTA (Chen et al., 2016) places SVs more accurately than other short-read callers, which may be a result of breakpoint assemblies MANTA performs.

Pangenomes

Pangenome graphs are constructed from multiple haplotypes and can be used to negate differences in alignments. The Pangenome Graph Builder (PGGB) (Garrison et al., 2023) constructs graphs from multiple haplotypes simultaneously, and the Minigraph-Cactus (MC) approach iteratively adds haplotypes to a graph (Hickey et al., 2022). Both were featured in the recent pangenome drafts constructed from 94 phased assemblies derived from 47 diverse samples recently released by the Human Pangenome Reference Consortium (HPRC) (Liao *et al.*, 2023).

Across unique loci, we identified all SVs that were present in more than one haplotype and matched an SV identified by MC (4,851 insertions, 3,240 deletions). We find that the MC

267 breakpoint offset is greater than all the HGSVC offsets for 69% of SV insertions and 41% of SV
268 deletions. For PGGB (4,831 insertions, 3,218 deletions), we find 13% of SV insertions and 15%
269 of SV deletions have a greater offset.

270 By manually inspecting many SVs, MC appears to place SV breakpoints irrespective of small
271 variants and does left-align against the reference. For example, a 2.1 kbp insertion was
272 identified in all HGSVC haplotypes (AF = 100%) with breakpoint variation, but it is shifted by 43
273 bp in the MC graph. This variant inserted into a TE and had TE sequence at the breakpoint
274 creating a tract of imperfect microhomology, and MC aligned 43 bp from the insertion sequence
275 to the reference. As a result, two false SNPs are found in all haplotypes with the SV and may
276 mislead downstream analyses. For example, SNPs linked to SVs do suggest mechanisms of SV
277 formation (Beck et al., 2019; Carvalho and Lupski, 2016; Deem et al., 2011), although no point
278 mutations were actually generated by this SV. (**Fig 4A**). This pattern was observed frequently in
279 the MC callset.

280 Many differences in the PGGB SVs are attributable to different breakpoint choices among
281 largely equivalent representations. For example, a 162 bp VNTR expansion (27 bp motif) with
282 one imperfect reference copy is inserted to the right of the reference copy rather than the left
283 (**Fig S3**). More importantly, we find a distinct pattern of PGGB deleting and re-inserting the
284 same bases when calling variants in loci without clean breakpoints. In one example, minimap2
285 represents a 101 bp net gain as a 109 bp insertion with three deletions totaling 8 bp, PGGB
286 calls a 118 bp insertion with a single 17 bp deletion, and MC calls a 105 bp insertion, a 5 bp
287 insertion, two deletions totaling 9 bp, and a SNP (**Fig 4B**). Further inspection of the breakpoints
288 shows that 13 bases deleted by PGGB are re-inserted as part of the SV insertion (**Fig 4C**). This
289 SV insertion sequence does not align to the human reference, but is present in the *Pan*
290 *troglodytes* on Chromosome 2 and is also in other primate genomes. Therefore, the insertion

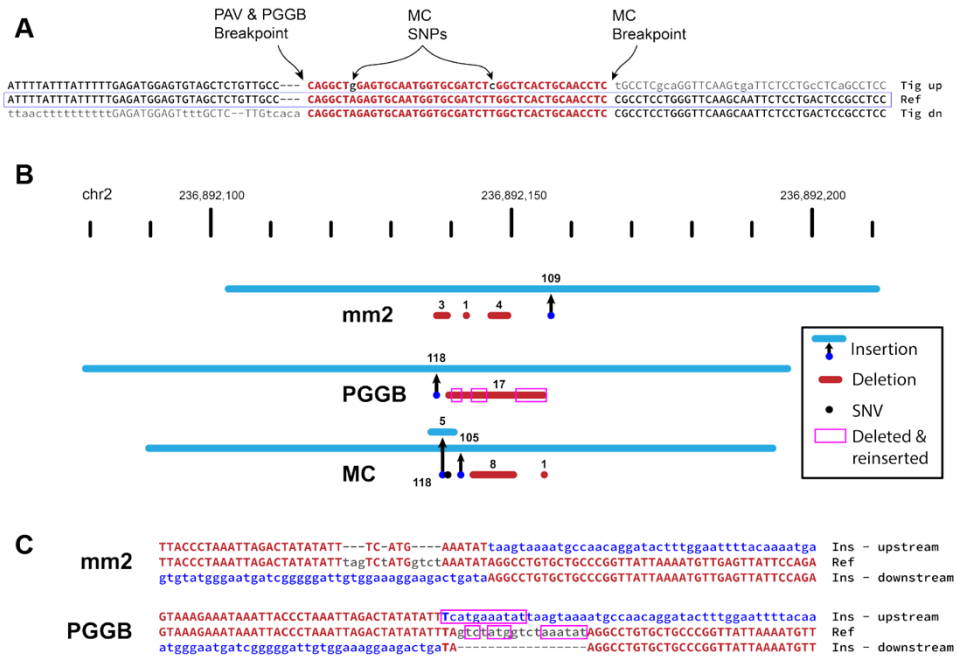

Figure 4: Pangenome graph breakpoints exhibit systematic bias. (A) A variant with a different breakpoint in the MC graph vs GRCh38 with the red portion showing the imperfect homology between breakpoint placements resulting in two SNPs called in the MC graph in all HPRC haplotypes (A>G and T>C). (**B**) An SV insertion (blue) paired with deleted bases (red) yields a 101 bp net gain by minimap2 (mm2, HGSVC callset), PGGB, and MC. minimap2 calls three small deletions near the insertion breakpoint. PGGB calls one larger deletion, but re-inserts deleted bases (magenta boxes) into the insertion call resulting in a larger SV insertion than minimap2. MC calls two insertions, two deletions, and a mismatch (black dot). (C) Alignments through breakpoints are shown for minimap2 and PGGB. Bases aligned to the reference are shown in red with matches in upper-case, inserted sequences are blue, and deleted bases are gray. The inserted sequence was not found in GRCh38, and likely represents the deletion of ancestral sequence, where the reference contains the derived (deleted) allele.

and is likely ancestral and the deletion became the reference allele by chance. The minimap2 representation of this locus appears to be the most likely biological explanation for this event with small template switches within the replication fork, which is characteristic of some repair mechanisms (Carvalho et al., 2013), most notably MMBIR (Hastings et al., 2009). Given that the insertion is ancestral, the deletion and re-insertion of bases is less likely.

296 In addition to creating different representations of SVs, the area between breakpoints is often
297 filled with small variants that are annotated differently across the haplotypes which may impact
298 the interpretation of variants. These different breakpoints intersect coding sequences for 26
299 genes on average in MC and 5 genes on average in PGGB with additional discrepancies in
300 UTRs and ncRNAs (**Table 2**). For example, we find a 180 bp insertion in *ESYT3* where
301 minimap2 and PGGB place the breakpoint in an intron, but MC places it in an exon (**Fig 5**).
302

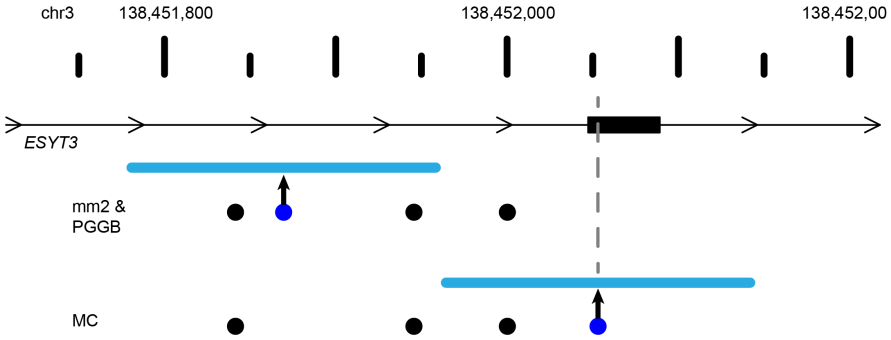

Figure 5: Breakpoint differences affect biological interpretation. A 180 bp insertion was called by minimap2 (mm2) and Pangenome Graph Builder (PGGB) at the same location within the intron of *ESYT3*, but Minigraph-Cactus (MC) placed the insertion 183 bp upstream inside an exon of *ESYT3*. SV insertions breakpoint locations are dark blue dots with the size of the SV shown as a light blue line. Point mutations are black dots. A gray line denotes the SV insertion location in the *ESYT3* exon.

Table 2: Consequences of differential breakpoints on RefSeq genes.

Graph	SV Type	Sample	CDS	5′ UTR	3′ UTR	ncRNA
mc	Insertions	HG03486	22	155	21	212
mc	Insertions	HG02818	19	168	15	182
mc	Deletions	HG03486	4	85	4	108
mc	Deletions	HG02818	6	78	4	113
pggb	Insertions	HG03486	12	53	8	74
pggb	Insertions	HG02818	12	55	7	69
pggb	Deletions	HG03486	3	41	2	47
pggb	Deletions	HG02818	2	34	1	36

Discussion

Advances in long-read sequencing coupled with new phased assembly and variant detection methods have increased the number of detectable SVs dramatically from less than 10,000 to more than 25,000 per diploid genome, and these advances continue to rival short-read technology by reducing costs, increasing availability, and improving read quality. In addition to detecting more SVs, long-reads also capture the full SV sequence, which is important for detailed analyses of non-reference sequences and has already proven to be transformative in mobile element characterization (Ebert *et al.*, 2021; Ferraj *et al.*, 2023).

Assemblies have improved breakpoint accuracy, however, systematic errors still exist where breakpoints span homologous regions. This effect is especially large for tandemly duplicated sequence and SV anchored in TEs. Since a majority of SVs have some form of homology around breakpoints (Ebert *et al.*, 2021; Lam *et al.*, 2010), the effect of differential breakpoints is potentially large even outside highly-repetitive sequences. However, modern aligners along with assembly-based SV detection consistently place SVs effectively reducing potential biases, but not eliminating them.

While errors in sequence and assembly do contribute to breakpoint differences, the most significant driver is the presence of small allelic changes near SV breakpoints, largely due to ancestral differences. This includes variation within insertions, which accumulate

polymorphisms over generations. Short-reads are subject to known reference biases where distant haplotypes align less confidently to alternate reference alleles (Brandt *et al.*, 2015; Degner *et al.*, 2009). Although small polymorphisms are spanned by much longer flanking sequences with long reads and assemblies, this reference bias now manifests as differential breakpoints.

On a large scale, these small differences have little effect on callset quality because modern variant merging and comparison tools do allow for imprecise breakpoints, however, it does impact breakpoint annotations. This impedes precise mechanistic inferences since shifting a breakpoint changes microhomology annotations by an average of more than 20 bp and leads to a lack of polymorphisms flanking SVs. These polymorphisms can be signatures of the DNA repair causing the rearrangement (Beck *et al.*, 2019; Carvalho and Lupski, 2016; Deem *et al.*, 2011). As a result, callsets are still imprecise and incomplete, even within unique loci, despite being covered by long, contiguous, high-quality assemblies.

While pangenome graphs normalize SV loci across samples, additional developments are required to improve breakpoint precision. PGGB agrees with minimap2 for many SVs; some method tuning could potentially improve PGGB for SV breakpoints in unique loci, whereas more substantial improvements may be required for MC. As graph methods mature, they hold promise for calling variants at scale across divergent haplotypes. Importantly, rare and somatic variants will not be in graph references based on population samples, and calling variants against the closest reference path will face the many of the same challenges as methods based on linear references. Improving breakpoints in graph representations and linear references will ultimately increase the utility of pangenome references.

In this study, we investigate breakpoint disagreements in unique regions of the human genome where long reads and assemblies spanning rearrangements with megabases of flanking sequence and few assembly errors. Complex genomic loci are dense with repeats and

346 breakpoint homology, and our results suggest that these loci present with larger breakpoint
347 discrepancies. While these loci have added complexity from larger, more frequent, and more
348 complex rearrangements as well as more collapsed reference loci (Sulovari et al., 2019; Vollger
349 et al., 2022), making more rigorous methods for precise rearrangement breakpoints may help
350 solve these regions more effectively.

351 Both simple and complex loci provide a rich opportunity for new methods to improve alignments,
352 variant calling, and variant annotation. While current sequencing data captures these events
353 with few errors, the limitations of current methods lead to systematic biases that affect the
354 accuracy of variant calls and limit their utility for detailed downstream analyses. While long
355 reads continue to gain in length and fidelity, the tools used to analyze them must keep pace.

356 Methods

357 Statistical analysis

358 Summary stats, such as mean and SD, were performed with Python numpy (v1.22.4) and
359 statistical tests including Student's t-test, Welch's t-test, F-test, and Fisher's exact test were
360 carried out with scipy (1.9.3). All tests were two-tailed. F-tests were used to determine if a
361 Student's t-test was carried out (F-test *p*-value ≥ 0.01) or a Welch's t-test (F-test *p*-value < 0.01).

362 *P*-values less than 1.0×10^{-100} are reported as $p < 1.0\times10^{-100}$. Extremely low *p*-values less than
363 the smallest floating point value Python can represent (~ $1\times10^{-308} \pm 1\times10^{-15}$ on our system) are
364 also reported as $p < 1\times10^{-100}$ in this manuscript.

365 *Microhomology.* The number of unique breakpoints was compared to the number of unique
366 microhomology calls per merged variant. Neither the number of unique breakpoint locations and
367 unique microhomology calls model a normal distribution ($p < 1\times10^{-100}$, scipy.stats.normaltest

based on D'Agostino and Pearson's test), so we computed correlation based on Spearman rank-order correlation coefficient.

Genome reference

We use the hg38-NoALT reference published with the HGSVC callset (Ebert *et al.*, 2021) (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/technical/reference/20200513_hg38_NoALT/). This reference is the full primary assembly of the human genome build 38 (GRCh38/hg38) including unplaced and unlocalized contigs, but it does not include patches, alternates, or decoys.

Ebert callset

We acquired the version 2 (Freeze 4) merged callset from HGSVC (Ebert *et al.*, 2021) (. We retained the same 32 population samples excluding the trio children used in the HGSVC publication. Frequencies and allele counts were adjusted to exclude child samples in the merged callset. We removed variants on unplaced and unlocalized contigs of the reference including only variants on primary chromosome scaffolds. A merging bug in SV-Pop allowed for some long-range intersects, and we removed these merged variants. To accomplish this filter this, we required either (a) the maximum offset is less than or equal to the merged SV length or (b) the maximum offset difference was less than 400 bp (200 bp in either direction) and the maximum SV length difference was not greater than 50% of the maximum SV length. These parameters mirror the expected results from the merging process without the long-range bug.

We obtained the Tandem Repeats Finder (TRF) (Benson, 1999) and RepeatMasker (Smit, 2013-2015) annotations from the UCSC Genome Browser (retrieved 2023-01-27, tracks "simpleRepeat" and "rmsk", respectively) (Kent et al., 2002). From TRF, we used all loci. From RMSK, we used all loci annotated as "Low_complexity" or "Simple_repeat". RMSK and TRF records within 200 bp were merged with BedTools merge (v2.30.0) (Quinlan and Hall, 2010)

and a 200 bp flank was added to all merged regions to each and intersected variants with both TRF and RMSK. Insertions were marked as tandem repeats if their insertion point was within a padded repeat region. Deletions were marked as tandem repeats if either reference breakpoint was within a 200 bp padded repeat region. Intersects were done with BedTools intersect (v2.30.0) (Quinlan and Hall, 2010).

Segmental duplications were annotated using the same process as tandem repeats. The segmental duplication track was retrieved from the UCSC genome browser (2023-01-28, track " genomicSuperDups"). Regions were merged within 200 bp and a 200 bp flank was added, both operations with BedTools. Insertion and deletion breakpoints were intersected with the merged and padded SD regions in the same way as repeats.

Distance between variant breakpoints is defined as before in SV-Pop (https://github.com/EichlerLab/svpop) as used by HGSVC for merging: min([start offset, end offset]) where "start offset" is the distance between the variant start positions and "end offset" is the distance between the variant end positions, which may be different if the variant is a deletion (Ebert *et al.*, 2021).

Pairwise comparisons were done by selecting all combinations of 64 haplotypes among the 32 samples (2,016 combinations of 2 haplotypes from a pool of 64). We obtained the original locations for each variant in all 64 haplotypes by tracing the merged call back through the sample to the original PAV call for each.

TEMR annotations.

We labeled SVs as TEMRs if TE annotations at SV breakpoints was consistent with a rearrangement mediated by TE homology. For reference repeats, we obtained the UCSC RepeatMasker (RMSK) track (database: rmsk.txt.gz) for hg38 (retrieved 2023-01-03). We retained only records with repeat class "LINE", "SINE", or "LTR" and with a minimum size of 100

bp. For deletions, we intersected the reference locations for each event independently (i.e. upstream breakpoint location and downstream breakpoint location) and annotated deletions as TEMRs if (a) both breakpoints intersected a TE annotation of the same type (e.g. Alu, ERV1, ERVK, L1, L2, etc), and (b) each side of the breakpoint intersected a different TE (i.e. distinct TE events). For SV insertions, we intersected the reference breakpoint with the RMSK track. We additionally obtained RepeatMasker annotations run on the merged callset by HGSVC (Ebert *et al.*, 2021) and selected repeat annotations within 10 bp of each end of the insertion. We annotated insertions as TEs if (a) RMSK annotations at each end of the inserted sequence and at the reference breakpoint were all the same TE type, and (b) RMSK annotation at each end of the insertion sequence were not the same TE (i.e. distinct TE events). Breakpoint intersections with the RMSK track were performed with BedTools intersect (v2.30.0).

Tandem duplication identification.

Insertion sequences in unique loci (excluding annotated SDs and TRs) were re-mapped to the reference with BLAST (v2.13.0) with parameters "-word_size 16 -perc_identity 95" against a BLAST database constructed from hg38-NoALT. We compiled a list of filter regions by including all TRs and RepeatMasker (RMSK) annotations with a score of 50 or greater from the UCSC browser and merging records with BedTools merge (v2.30.0). BLAST alignments were discarded if 50% or more of the alignment record intersected the TR and RMSK filter. We further filtered BLAST hits to include only records that mapped within 10% of the SV length from the insertion site or deletion breakpoints (e.g. 1 kbp INS, 100 bp window around the insertion site) with a minimum of 100 bp for small SVs. For deletions, we removed the deletion sequence alignment (i.e. remapping produces an alignment over the deletion). Alignments less than 30 bp were also excluded. Some redundant overlapping alignments remained and appeared to be driven by small TRs that were not in the reference, which were removed by keeping only the longest record if records overlapped by 80% or more. The same 80% overlap was performed in

both reference space using aligned reference coordinates and in SV sequence space using coordinates from the SV sequence (i.e. the first base of the SV sequence is position 0). We selected SVs where the total number of aligned bases on each side of the breakpoint was within 90% of the total SV size and ensuring records with large gaps spanning more bases than were aligned did not contribute to the SV size calculation. We did not select records that had the expected alignment pattern (i.e. upstream SV sequence mapping downstream of the SV breakpoint and downstream SV sequence mapping upstream of the SV), although all the records left after the filtering process did exhibit this pattern.

Small variants around breakpoints

Our goal was to identify small differences between haplotypes that causes variant breakpoints to be placed differently. For each haplotype pair, we selected SV insertions and deletions with breakpoints placed at different sites and with breakpoints in unique loci (not TR or SD). We extracted the haplotype sequence from around the assembly including a 50 bp flank on each side and we extended one end appropriately to add flank so that in the absence of other small variants, both sequences should start on the same base.

The sequences were aligned so that the right-most variant was the reference and the left-most variant was the query, although either order should produce similar results. Sequences were aligned with the "swalign" Python package (v0.3.7) using a global alignment and with match, mismatch, and gap scores from the minimap2 (short-gap scores from the default double-affine parameters):

```
aligner = swalign.LocalAlignment(
   swalign.NucleotideScoringMatrix(2, -4),
   gap_penalty=-4, gap_extension_penalty=-2,
   globalalign=True
```

Alignment align ("M" CIGAR operations) records were transformed to match/mismatch ("=" and "X" CIGAR operations), and using the known flanks added to each, we assigned variants to left flank, left breakpoint (intersecting the breakpoint), differential region, right breakpoint (intersecting the breakpoint), and right flank along with their relative position in each category.

Microhomology.

Microhomology is the span of perfectly matching bases at each end of a breakpoint, for example, the perfect homology at sites of ectopic recombination (i.e., NAHR), homology directed repair, replication-based repair, or alt-EJ. We measured homology at breakpoints using an algorithm in PAV and previously validated as part of a TEMR project (Balachandran *et al.*, 2022) where the region upstream of an SV sequence is matched with the downstream reference or contig, and the region downstream of the SV sequence is matched with the upstream reference or contig. To compare haplotypes more consistently, we computed SV homology for insertions against the upstream and downstream contig where the SV was called, and against the reference for deletions. We excluded all TD variants from homology because estimating breakpoint homology using this method counts whole TD copies as homologous.

Graph genome comparisons.

Variants against GRCh38 for PGGB and MC graphs were obtained from the decomposed VCFs published by the HPRC (Liao *et al.*, 2023) (see "Resource table" in Methods for URLs). Variants were extracted for each sample using SV-Pop (Ebert *et al.*, 2021). Differences were manually investigated using the UCSC browser and custom browser tracks for HGSVC and HPRC variants.

Resource table

Resource	Source	Identifier / version
Merged Ebert callset	Ebert *et al.* (2021) https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/release/v2.0/integrated_callset/	v2.0 (Freeze4)
Python		3.9.15
Numpy		1.22.4
Scipy		1.9.3
BedTools		2.30.0
BLAST		2.13.0
swalign	Python package	0.3.7
HPRC PGGB decomposed VCF	https://s3-us-west-2.amazonaws.com/human-pangenomics/pangenomes/freeze/freeze1/pggb/vcfs/hprc-v1.0-pggb.grch38.vcfbub.a100k.wave.vcf.gz	
HPRC MC decomposed VCF	https://s3-us-west-2.amazonaws.com/human-pangenomics/pangenomes/freeze/freeze1/minigraph-cactus/hprc-v1.0-mc.grch38.vcfbub.a100k.wave.vcf.gz	

Competing Interests Statement

The authors have no competing interests to disclose.

Acknowledgments

P.A.A and C.R.B were supported by NIH NIGMS R35GM133600 and NIH NCI P30CA034196. The Human Genome Structural Variation Consortium (HGSVC) provided published data, support, and feedback, and the HGSVC was supported by NIH NHGRI U24HG007497. Thanks to Parithi Balachandran for helping to check duplication detection tools and SV breakpoints. Thanks to Evan E. Eichler for providing feedback on graph genome comparisons.

References

Abel, H.J., Larson, D.E., Regier, A.A., Chiang, C., Das, I., Kanchi, K.L., Layer, R.M., Neale, B.M., Salerno, W.J., Reeves, C., et al. (2020). Mapping and characterization of structural variation in 17,795 human genomes. Nature *583*, 83-89. 10.1038/s41586-020-2371-0.

Alkan, C., Coe, B.P., and Eichler, E.E. (2011). Genome structural variation discovery and genotyping. Nat Rev Genet *12*, 363-376. 10.1038/nrg2958.

Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L., Suresh, H., Ramakrishnan, S., Maumus, F., Ciren, D., et al. (2020). Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. Cell *182*, 145-161 e123. 10.1016/j.cell.2020.05.021.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J Mol Biol *215*, 403-410. 10.1016/S0022-2836(05)80360-2.

Arlt, M.F., Mulle, J.G., Schaibley, V.M., Ragland, R.L., Durkin, S.G., Warren, S.T., and Glover, T.W. (2009). Replication stress induces genome-wide copy number changes in human cells that resemble polymorphic and pathogenic variants. Am J Hum Genet *84*, 339-350. 10.1016/j.ajhg.2009.01.024.

Audano, P.A., Sulovari, A., Graves-Lindsay, T.A., Cantsilieris, S., Sorensen, M., Welch, A.E., Dougherty, M.L., Nelson, B.J., Shah, A., Dutcher, S.K., et al. (2019). Characterizing the Major Structural Variant Alleles of the Human Genome. Cell *176*, 663-675 e619. 10.1016/j.cell.2018.12.019.

Bailey, J.A., Gu, Z., Clark, R.A., Reinert, K., Samonte, R.V., Schwartz, S., Adams, M.D., Myers, E.W., Li, P.W., and Eichler, E.E. (2002). Recent segmental duplications in the human genome. Science *297*, 1003-1007. 10.1126/science.1072047.

Balachandran, P., Walawalkar, I.A., Flores, J.I., Dayton, J.N., Audano, P.A., and Beck, C.R. (2022). Transposable element-mediated rearrangements are prevalent in human genomes. Nat Commun *13*, 7115. 10.1038/s41467-022-34810-8.

Beck, C.R., Carvalho, C.M., Banser, L., Gambin, T., Stubbolo, D., Yuan, B., Sperle, K., McCahan, S.M., Henneke, M., Seeman, P., et al. (2015). Complex genomic rearrangements at

the PLP1 locus include triplication and quadruplication. PLoS Genet *11*, e1005050. 10.1371/journal.pgen.1005050.

Beck, C.R., Carvalho, C.M.B., Akdemir, Z.C., Sedlazeck, F.J., Song, X., Meng, Q., Hu, J., Doddapaneni, H., Chong, Z., Chen, E.S., et al. (2019). Megabase Length Hypermutation Accompanies Human Structural Variation at 17p11.2. Cell *176*, 1310-1324 e1310. 10.1016/j.cell.2019.01.045.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res *27*, 573-580. 10.1093/nar/27.2.573.

Beyter, D., Ingimundardottir, H., Oddsson, A., Eggertsson, H.P., Bjornsson, E., Jonsson, H., Atlason, B.A., Kristmundsdottir, S., Mehringer, S., Hardarson, M.T., et al. (2021). Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. Nat Genet *53*, 779-786. 10.1038/s41588-021-00865-4.

Brandt, D.Y., Aguiar, V.R., Bitarello, B.D., Nunes, K., Goudet, J., and Meyer, D. (2015). Mapping Bias Overestimates Reference Allele Frequencies at the HLA Genes in the 1000 Genomes Project Phase I Data. G3 (Bethesda) *5*, 931-941. 10.1534/g3.114.015784.

Carvalho, C.M., and Lupski, J.R. (2016). Mechanisms underlying structural variant formation in genomic disorders. Nat Rev Genet *17*, 224-238. 10.1038/nrg.2015.25.

Carvalho, C.M., Pehlivan, D., Ramocki, M.B., Fang, P., Alleva, B., Franco, L.M., Belmont, J.W., Hastings, P.J., and Lupski, J.R. (2013). Replicative mechanisms for CNV formation are error prone. Nat Genet *45*, 1319-1326. 10.1038/ng.2768.

Carvalho, C.M., Ramocki, M.B., Pehlivan, D., Franco, L.M., Gonzaga-Jauregui, C., Fang, P., McCall, A., Pivnick, E.K., Hines-Dowell, S., Seaver, L.H., et al. (2011). Inverted genomic segments and complex triplication rearrangements are mediated by inverted repeats in the human genome. Nat Genet *43*, 1074-1081. 10.1038/ng.944.

Chaisson, M.J., Huddleston, J., Dennis, M.Y., Sudmant, P.H., Malig, M., Hormozdiari, F., Antonacci, F., Surti, U., Sandstrom, R., Boitano, M., et al. (2015). Resolving the complexity of the human genome using single-molecule sequencing. Nature *517*, 608-611. 10.1038/nature13907.

Chaisson, M.J.P., Sanders, A.D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E.J., Rodriguez, O.L., Guo, L., Collins, R.L., et al. (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat Commun *10*, 1784. 10.1038/s41467-018-08148-z.

Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Kallberg, M., Cox, A.J., Kruglyak, S., and Saunders, C.T. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics *32*, 1220-1222. 10.1093/bioinformatics/btv710.

Collins, R.L., Brand, H., Karczewski, K.J., Zhao, X., Alfoldi, J., Francioli, L.C., Khera, A.V., Lowther, C., Gauthier, L.D., Wang, H., et al. (2020). A structural variation reference for medical and population genetics. Nature *581*, 444-451. 10.1038/s41586-020-2287-8.

Deem, A., Keszthelyi, A., Blackgrove, T., Vayl, A., Coffey, B., Mathur, R., Chabes, A., and Malkova, A. (2011). Break-induced replication is highly inaccurate. PLoS Biol *9*, e1000594. 10.1371/journal.pbio.1000594.

Degner, J.F., Marioni, J.C., Pai, A.A., Pickrell, J.K., Nkadori, E., Gilad, Y., and Pritchard, J.K. (2009). Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. Bioinformatics *25*, 3207-3212. 10.1093/bioinformatics/btp579.

Ebert, P., Audano, P.A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., Bonder, M.J., Sulovari, A., Ebler, J., Zhou, W., Serra Mari, R., et al. (2021). Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science *372*. 10.1126/science.abf7117.

Ebler, J., Ebert, P., Clarke, W.E., Rausch, T., Audano, P.A., Houwaart, T., Mao, Y., Korbel, J.O., Eichler, E.E., Zody, M.C., et al. (2022). Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. Nat Genet *54*, 518-525. 10.1038/s41588-022-01043-w.

Eizenga, J.M., Novak, A.M., Sibbesen, J.A., Heumos, S., Ghaffaari, A., Hickey, G., Chang, X., Seaman, J.D., Rounthwaite, R., Ebler, J., et al. (2020). Pangenome Graphs. Annu Rev Genomics Hum Genet *21*, 139-162. 10.1146/annurev-genom-120219-080406.

Encode Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74. 10.1038/nature11247.

Encode Project Consortium, Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shoresh, N., Adrian, J., Kawli, T., Davis, C.A., Dobin, A., et al. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature *583*, 699-710. 10.1038/s41586-020-2493-4.

English, A.C., Menon, V.K., Gibbs, R.A., Metcalf, G.A., and Sedlazeck, F.J. (2022). Truvari: refined structural variant comparison preserves allelic diversity. Genome Biol *23*, 271. 10.1186/s13059-022-02840-6.

Ferraj, A., Audano, P.A., Balachandran, P., Czechanski, A., Flores, J.I., Radecki, A.A., Mosur, V., Gordon, D.S., Walawalkar, I.A., Eichler, E.E., et al. (2023). Resolution of structural variation in diverse mouse genomes reveals chromatin remodeling due to transposable elements. Cell Genomics. 10.1016/j.xgen.2023.100291.

Frankish, A., Diekhans, M., Jungreis, I., Lagarde, J., Loveland, J.E., Mudge, J.M., Sisu, C., Wright, J.C., Armstrong, J., Barnes, I., et al. (2021). Gencode 2021. Nucleic Acids Res *49*, D916-D923. 10.1093/nar/gkaa1087.

Garg, S., Fungtammasan, A., Carroll, A., Chou, M., Schmitt, A., Zhou, X., Mac, S., Peluso, P., Hatas, E., Ghurye, J., et al. (2021). Chromosome-scale, haplotype-resolved assembly of human genomes. Nat Biotechnol *39*, 309-312. 10.1038/s41587-020-0711-0.

Garrison, E., Guarracino, A., Heumos, S., Villani, F., Bao, Z., Tattini, L., Hagmann, J., Vorbrugg, S., Marco-Sola, S., Kubica, C., et al. (2023). Building pangenome graphs. bioRxiv.

Han, K., Lee, J., Meyer, T.J., Remedios, P., Goodwin, L., and Batzer, M.A. (2008). L1 recombination-associated deletions generate human genomic variation. Proc Natl Acad Sci U S A *105*, 19366-19371. 10.1073/pnas.0807866105.

Hastings, P.J., Ira, G., and Lupski, J.R. (2009). A microhomology-mediated break-induced replication model for the origin of human copy number variation. PLoS Genet *5*, e1000327. 10.1371/journal.pgen.1000327.

Hickey, G., Monlong, J., Novak, A., Eizenga, J.M., Consortium, H.P.R., Li, H., and Paten, B. (2022). Pangenome Graph Construction from Genome Alignment with Minigraph-Cactus. bioRxiv.

ICGC TCGA Pan-Cancer Analysis of Whole Genomes Consortium (2020). Pan-cancer analysis of whole genomes. Nature *578*, 82-93. 10.1038/s41586-020-1969-6.

Iliakis, G., Murmann, T., and Soni, A. (2015). Alternative end-joining repair pathways are the ultimate backup for abrogated classical non-homologous end-joining and homologous

recombination repair: Implications for the formation of chromosome translocations. Mutat Res Genet Toxicol Environ Mutagen *793*, 166-175. 10.1016/j.mrgentox.2015.07.001.

International HapMap, C., Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., et al. (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851-861. 10.1038/nature06258.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921. 10.1038/35057062.

Jebb, D., Huang, Z., Pippel, M., Hughes, G.M., Lavrichenko, K., Devanna, P., Winkler, S., Jermiin, L.S., Skirmuntt, E.C., Katzourakis, A., et al. (2020). Six reference-quality genomes reveal evolution of bat adaptations. Nature *583*, 578-584. 10.1038/s41586-020-2486-3.

Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alfoldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434-443. 10.1038/s41586-020-2308-7.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res *12*, 996-1006. 10.1101/gr.229102.

Kirsche, M., Prabhu, G., Sherman, R., Ni, B., Aganezov, S., and Schatz, M.C. (2021). Jasmine: Population-scale structural variant comparison and analysis. bioRxiv.

Lam, H.Y., Mu, X.J., Stutz, A.M., Tanzer, A., Cayting, P.D., Snyder, M., Kim, P.M., Korbel, J.O., and Gerstein, M.B. (2010). Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. Nat Biotechnol *28*, 47-55. 10.1038/nbt.1600.

Lee, J.A., Carvalho, C.M., and Lupski, J.R. (2007). A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. Cell *131*, 1235-1247. 10.1016/j.cell.2007.11.037.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094-3100. 10.1093/bioinformatics/bty191.

Li, R., Gong, M., Zhang, X., Wang, F., Liu, Z., Zhang, L., Yang, Q., Xu, Y., Xu, M., Zhang, H., et al. (2023). A sheep pangenome reveals the spectrum of structural variations and their effects on tail phenotypes. Genome Research *33*, 463-477. 10.1101/gr.277372.122.

Li, Y., Roberts, N.D., Wala, J.A., Shapira, O., Schumacher, S.E., Kumar, K., Khurana, E., Waszak, S., Korbel, J.O., Haber, J.E., et al. (2020). Patterns of somatic structural variation in human cancer genomes. Nature *578*, 112-121. 10.1038/s41586-019-1913-9.

Liao, W.W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J.K., Monlong, J., Abel, H.J., et al. (2023). A draft human pangenome reference. Nature *617*, 312-324. 10.1038/s41586-023-05896-x.

Mao, Y., Catacchio, C.R., Hillier, L.W., Porubsky, D., Li, R., Sulovari, A., Fernandes, J.D., Montinaro, F., Gordon, D.S., Storer, J.M., et al. (2021). A high-quality bonobo genome refines the analysis of hominid evolution. Nature *594*, 77-81. 10.1038/s41586-021-03519-x.

Menghi, F., Inaki, K., Woo, X., Kumar, P.A., Grzeda, K.R., Malhotra, A., Yadav, V., Kim, H., Marquez, E.J., Ucar, D., et al. (2016). The tandem duplicator phenotype as a distinct genomic configuration in cancer. Proc Natl Acad Sci U S A *113*, E2373-2382. 10.1073/pnas.1520010113.

Morales, M.E., White, T.B., Streva, V.A., DeFreece, C.B., Hedges, D.J., and Deininger, P.L. (2015). The contribution of alu elements to mutagenic DNA double-strand break repair. PLoS Genet *11*, e1005016. 10.1371/journal.pgen.1005016.

Mouse Genome Sequencing Consortium, Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520-562. 10.1038/nature01262.

Nattestad, M., Goodwin, S., Ng, K., Baslan, T., Sedlazeck, F.J., Rescheneder, P., Garvin, T., Fang, H., Gurtowski, J., Hutton, E., et al. (2018). Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line. Genome Res *28*, 1126-1135. 10.1101/gr.231100.117.

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A.V., Mikheenko, A., Vollger, M.R., Altemose, N., Uralsky, L., Gershman, A., et al. (2022). The complete sequence of a human genome. Science *376*, 44-53. 10.1126/science.abj6987.

O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res *44*, D733-745. 10.1093/nar/gkv1189.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842. 10.1093/bioinformatics/btq033.

Rausch, T., Snajder, R., Leger, A., Simovic, M., Giurgiu, M., Villacorta, L., Henssen, A.G., Frohling, S., Stegle, O., Birney, E., et al. (2023). Long-read sequencing of diagnosis and post-therapy medulloblastoma reveals complex rearrangement patterns and epigenetic signatures. Cell Genom *3*, 100281. 10.1016/j.xgen.2023.100281.

Rozowsky, J., Gao, J., Borsari, B., Yang, Y.T., Galeev, T., Gursoy, G., Epstein, C.B., Xiong, K., Xu, J., Li, T., et al. (2023). The EN-TEx resource of multi-tissue personal epigenomes & variant-impact models. Cell *186*, 1493-1511 e1440. 10.1016/j.cell.2023.02.018.

Sakamoto, Y., Xu, L., Seki, M., Yokoyama, T.T., Kasahara, M., Kashima, Y., Ohashi, A., Shimada, Y., Motoi, N., Tsuchihara, K., et al. (2020). Long-read sequencing for non-small-cell lung cancer genomes. Genome Res *30*, 1243-1257. 10.1101/gr.261941.120.

Schneider, V.A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.C., Kitts, P.A., Murphy, T.D., Pruitt, K.D., Thibaud-Nissen, F., Albracht, D., et al. (2017). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res *27*, 849-864. 10.1101/gr.213611.116.

Sen, S.K., Han, K., Wang, J., Lee, J., Wang, H., Callinan, P.A., Dyer, M., Cordaux, R., Liang, P., and Batzer, M.A. (2006). Human genomic deletions mediated by recombination between Alu elements. Am J Hum Genet *79*, 41-53. 10.1086/504600.

Smit, A.F.H., R; Green, P. (2013-2015). RepeatMasker Open-4.0. http://www.repeatmasker.org.

Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M.H., et al. (2015). An integrated map of structural variation in 2,504 human genomes. Nature *526*, 75-81. 10.1038/nature15394.

Sulovari, A., Li, R., Audano, P.A., Porubsky, D., Vollger, M.R., Logsdon, G.A., Human Genome Structural Variation, C., Warren, W.C., Pollen, A.A., Chaisson, M.J.P., and Eichler, E.E. (2019). Human-specific tandem repeat expansion and differential gene expression during primate evolution. Proc Natl Acad Sci U S A *116*, 23243-23253. 10.1073/pnas.1912175116.

Taliun, D., Harris, D.N., Kessler, M.D., Carlson, J., Szpiech, Z.A., Torres, R., Taliun, S.A.G., Corvelo, A., Gogarten, S.M., Kang, H.M., et al. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature *590*, 290-299. 10.1038/s41586-021-03205-y.

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68-74. 10.1038/nature15393.

Turner, T.N., Coe, B.P., Dickel, D.E., Hoekzema, K., Nelson, B.J., Zody, M.C., Kronenberg, Z.N., Hormozdiari, F., Raja, A., Pennacchio, L.A., et al. (2017). Genomic Patterns of De Novo Mutation in Simplex Autism. Cell *171*, 710-722 e712. 10.1016/j.cell.2017.08.047.

Vogt, J., Bengesser, K., Claes, K.B., Wimmer, K., Mautner, V.F., van Minkelen, R., Legius, E., Brems, H., Upadhyaya, M., Hogel, J., et al. (2014). SVA retrotransposon insertion-associated deletion represents a novel mutational mechanism underlying large genomic copy number changes with non-recurrent breakpoints. Genome Biol *15*, R80. 10.1186/gb-2014-15-6-r80.

Vollger, M.R., Guitart, X., Dishuck, P.C., Mercuri, L., Harvey, W.T., Gershman, A., Diekhans, M., Sulovari, A., Munson, K.M., Lewis, A.P., et al. (2022). Segmental duplications and their variation in a complete human genome. Science *376*, eabj6965. 10.1126/science.abj6965.

Wahlster, L., Verboon, J.M., Ludwig, L.S., Black, S.C., Luo, W., Garg, K., Voit, R.A., Collins, R.L., Garimella, K., Costello, M., et al. (2021). Familial thrombocytopenia due to a complex structural variant resulting in a WAC-ANKRD26 fusion transcript. J Exp Med *218*. 10.1084/jem.20210444.

Willis, N.A., Frock, R.L., Menghi, F., Duffey, E.E., Panday, A., Camacho, V., Hasty, E.P., Liu, E.T., Alt, F.W., and Scully, R. (2017). Mechanism of tandem duplication formation in BRCA1-mutant cells. Nature *551*, 590-595. 10.1038/nature24477.

Zhou, W., Emery, S.B., Flasch, D.A., Wang, Y., Kwan, K.Y., Kidd, J.M., Moran, J.V., and Mills, R.E. (2020). Identification and characterization of occult human-specific LINE-1 insertions using long-read sequencing technology. Nucleic Acids Res *48*, 1146-1163. 10.1093/nar/gkz1173.

497

498

Supplemental Material

Figure S1

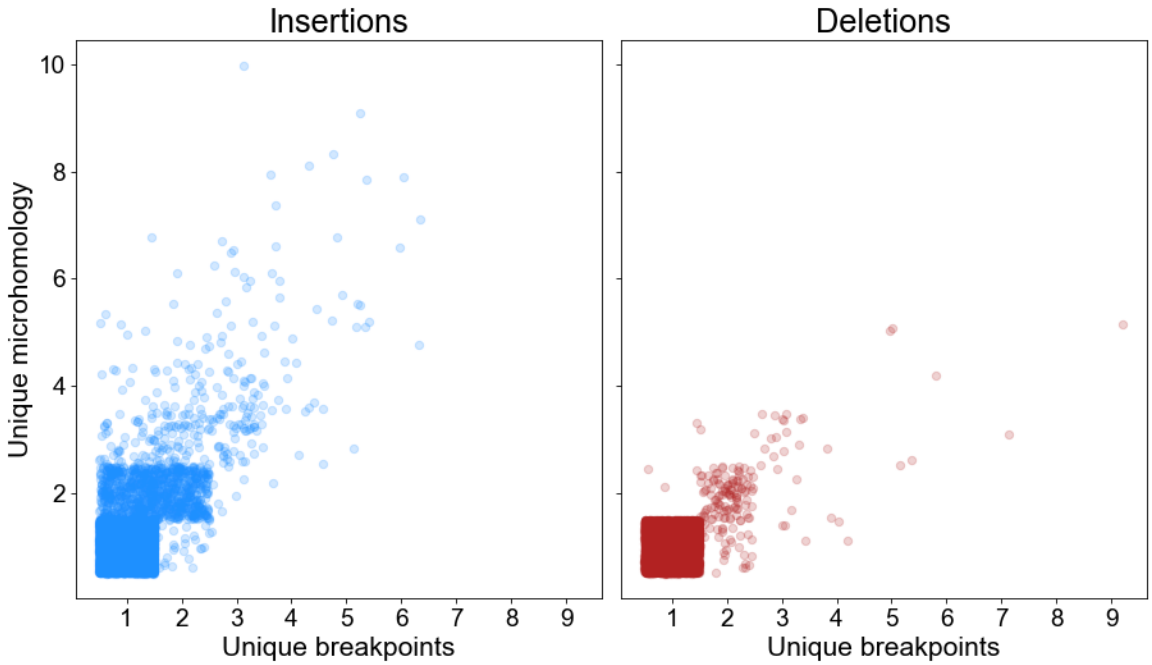

Figure S1: Breakpoint changes alter microhomology. The number of unique breakpoints (horizontal axis) a variant has across haplotypes has a dramatic impact on the number of unique microhomology annotations (violet line: best-fit least-squares regression line). Transparency and jittering (± 0.5) separates points falling on integers.

503 Figure S2

504

505

506

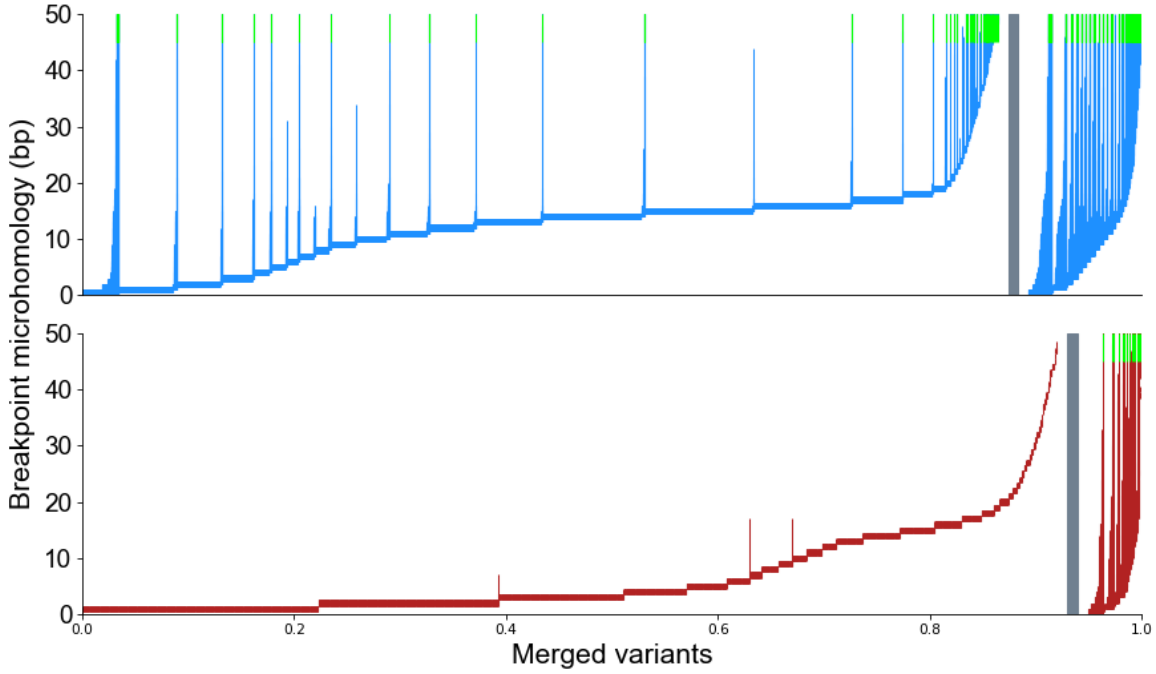

Figure S2: Breakpoint placement leads to large microhomology differences. For each merged SVs (insertions top/blue, deletions bottom/red), vertical lines extend from the minimum microhomology to the maximum microhomology across haplotypes. A gray bar separates SVs with consistent breakpoints (left) from SVs called at different breakpoints across haplotypes (right). Green tips denote lines that extend past the top of the figure.

Figure S3

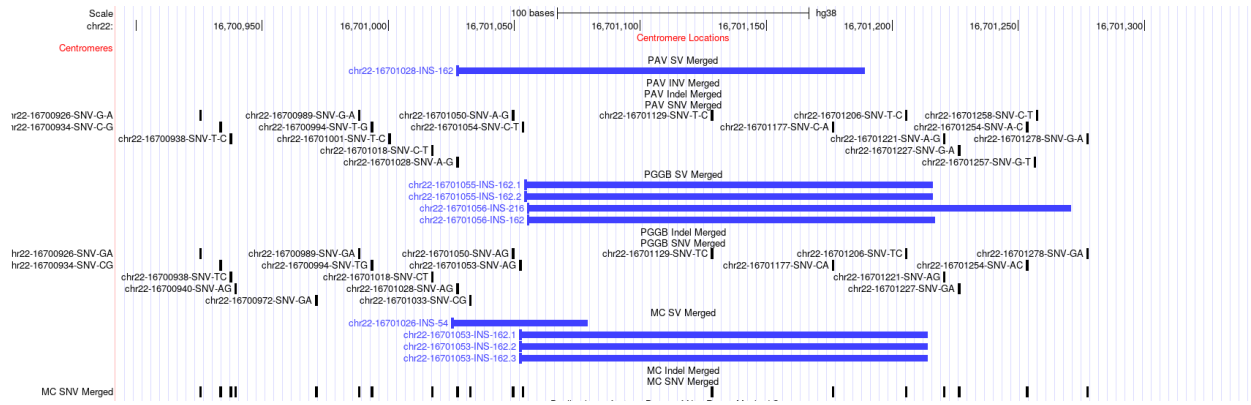

Figure S3: Ambiguous breakpoints for SVs in degenerate tandem repeats. The true breakpoint for this 162 bp expansion is difficult to identify even though tandem repeats in this locus were too diverged or too small to yield a tandem annotation. Despite this divergence, breakpoints were still not consistently placed, and the optimal location is difficult to identify and all three methods chose different breakpoints.

Table S1

Table S1. Offsets per haplotype pair. Average effect of shifted breakpoints on each pair of haplotypes (n = 2,016 combinations of 64 haplotypes). TR: Tandem Repeat, SD: Segmental Duplication, N: Number of variants, Diff: Variants with different offsets in the pair of haplotypes.

	Insertions			Deletions		
	N	Diff	Diff %	N	Diff	Diff %
No TR/SD	1,235	54	4.4%	391	7	1.7%
SD, no TR	58	6	9.8%	22	2	8.8%
All	4,862	967	20.0%	2,614	382	14.7%