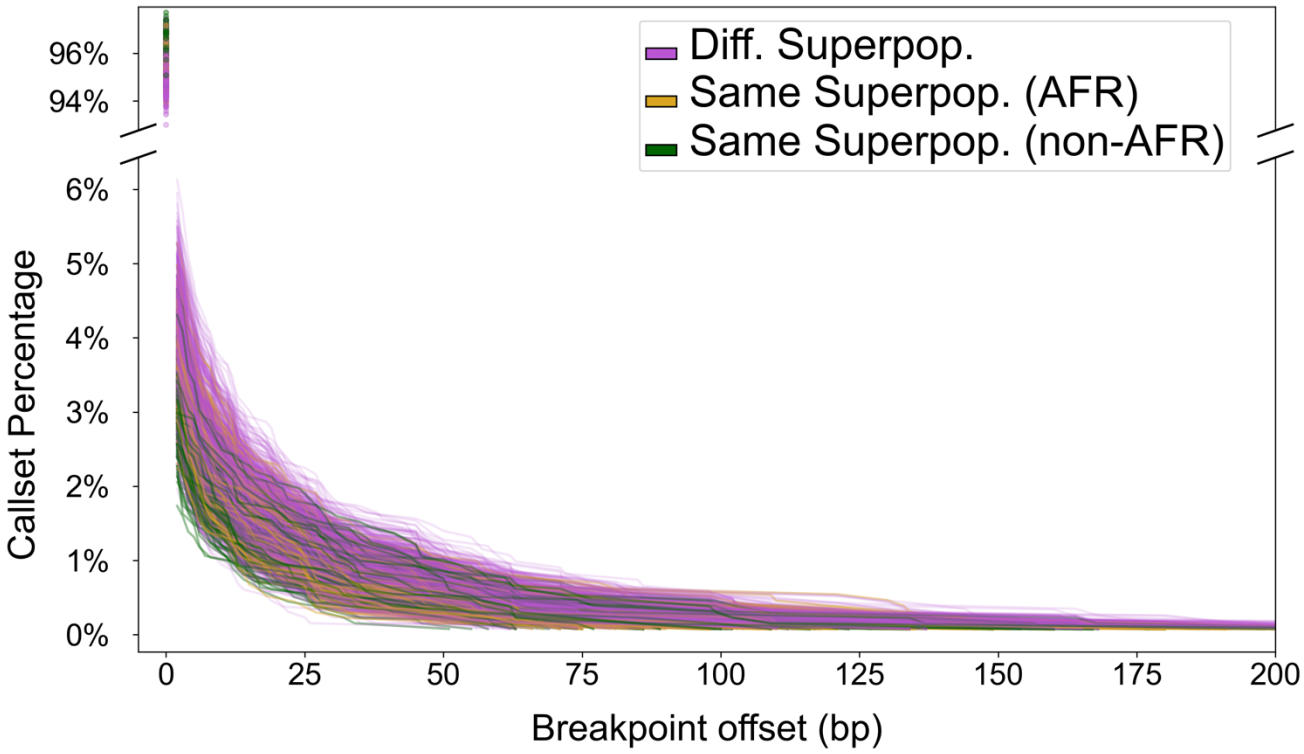
Ancestral bias alters structural variant breakpoints and hides true polymorphisms in euchromatin

Supplemental Figures

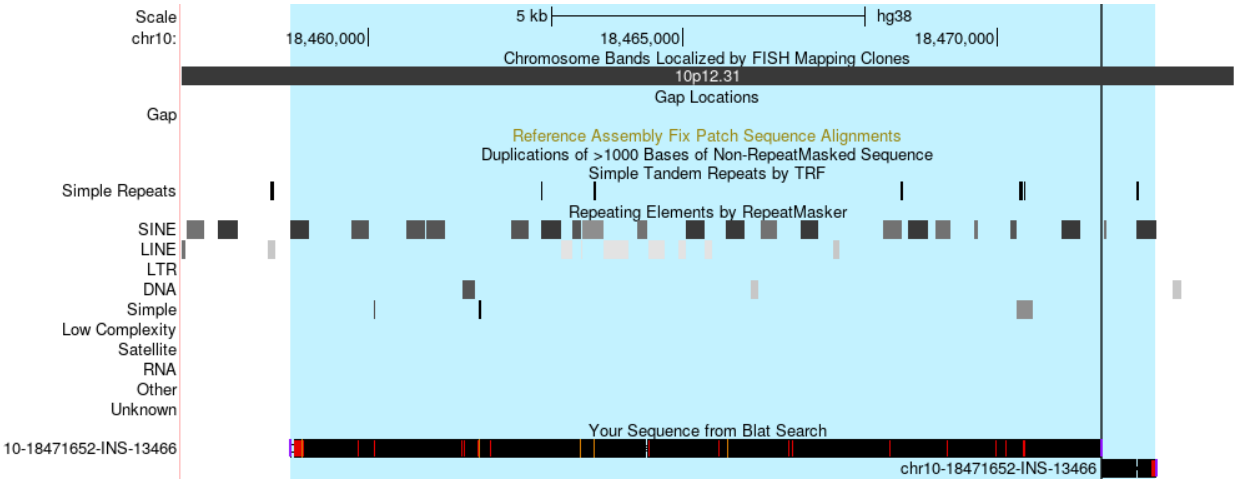
Contents

Supplemental Figure 1



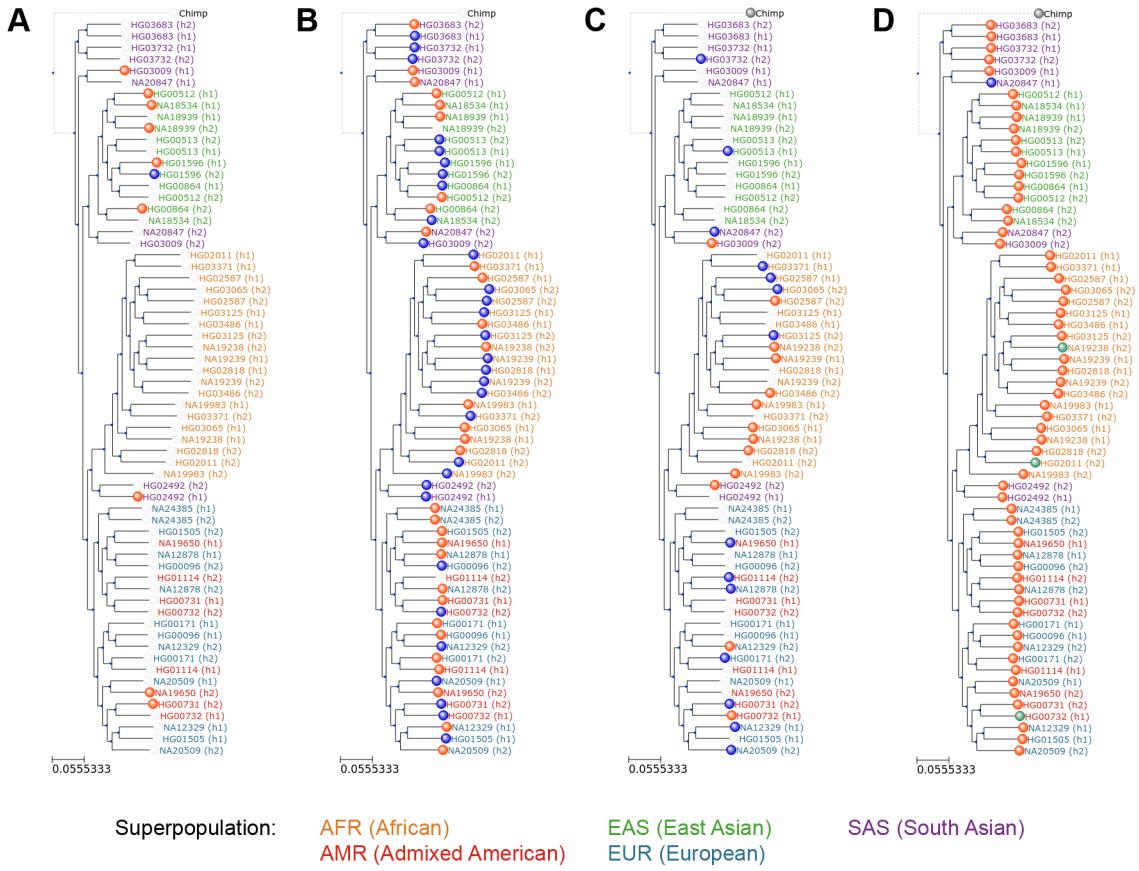
Supplemental Figure 1. A cumulative distribution of breakpoint offset for all haplotype pairs. Most variants in both haplotypes share the same breakpoint (upper y-axis). For variants with at least 1 bp offset (lower y-axis), the cumulative proportion of matched calls decreases with increasing breakpoint 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.

Supplemental Figure 2



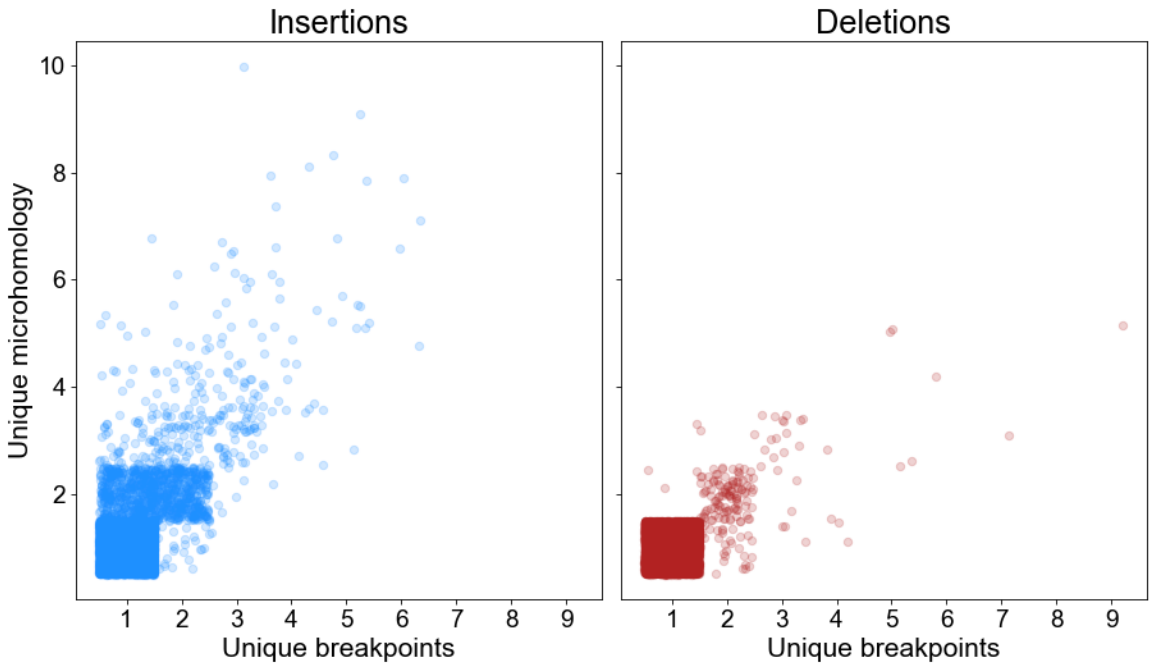
Supplemental Figure 2. SV breakpoint volatility for tandem duplications. A 13.5 kbp insertion was annotated as a tandem duplication by mapping the chimeric insertion sequence to the reference (blue) with BLAST where the alignment was split at the insertion sequence (black vertical line), which was reproduced by BLAT (bottom track). This SV breakpoint might have been produced by a duplication, in which case the breakpoint should be in a flanking *Alu*Sx element (black boxes in the SINE row at the edge of the blue region). Alternatively, it may represent a deletion event where the reference is the derived deleted allele and the sample is not deleted. Both possibilities can produce this SV representation, and more detailed analysis is needed to discern them.

Supplemental Figure 3



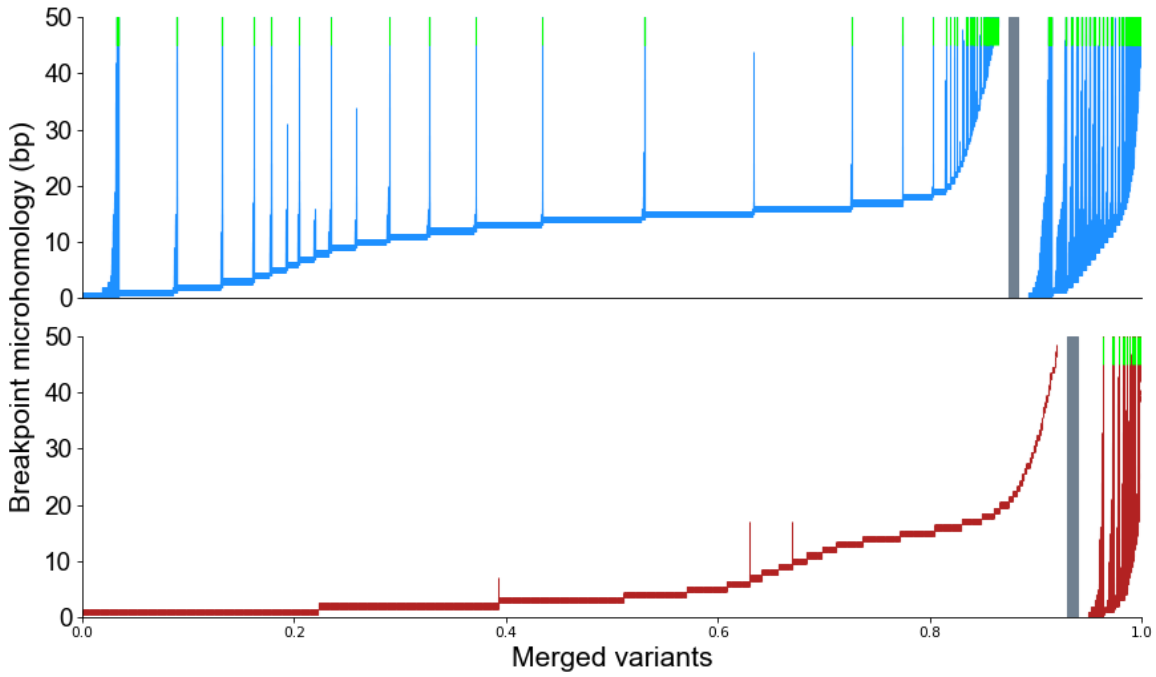
Supplemental Figure 3. Breakpoint differences in phylogenetic context. Four examples of SVs with different breakpoints in phylogenetic context across the 64 haplotypes with sample names colored by assigned 1000 Genomes superpopulations. Bubbles indicate SV was called in that haplotype and each color indicates a breakpoint location. A gray bubble indicates the SV was found in the Chimpanzee genome. To aid visualization, the Chimpanzee (outgroup) is displayed a shortened branch length (dashed line). (**A**) A 9 kbp insertion is detected in 10 diverse haplotypes with two closely-related haplotypes disagreeing on the breakpoint location. This SV may be recurrent, however, the breakpoint locations do not follow recurrent patterns. (**B**) A 5 kbp insertion detected in most haplotypes likely arose in early humans. A mix of breakpoint locations is found both in both African and non-African genomes at different locations. (**C**) A 3 kbp insertion was also found in the Chimpanzee genome is likely an ancestral deletion where the deleted allele became part of the reference genome, GRCh38. Two breakpoints for the non-deleted ancestral state show population stratification and are not likely related to SV biology. (**D**) An insertion present in all haplotypes and chimpanzee is a likely reference error or a very rare deletion that became reference. As in (C), different breakpoint locations for the non-deleted state are unrelated to SV formation.

Supplemental Figure 4



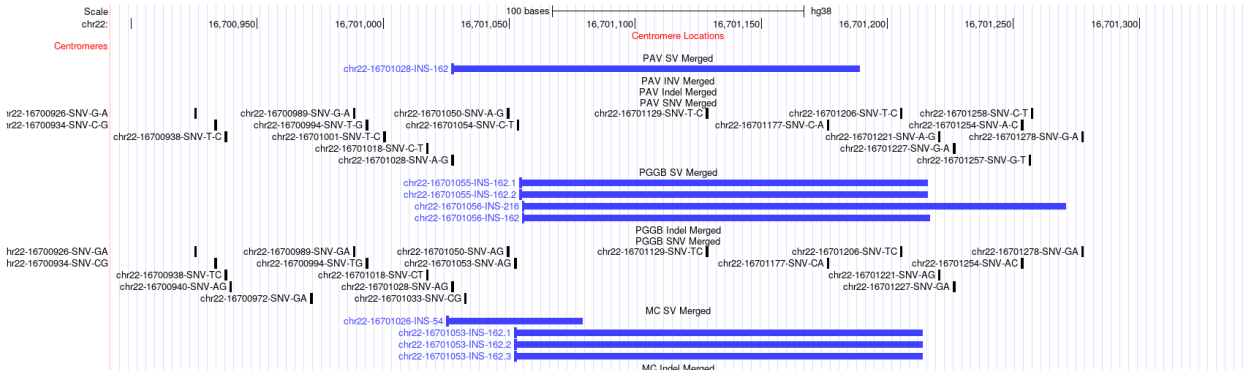
Supplemental Figure 4. Breakpoint changes alter microhomology. The number of unique breakpoints (horizontal axis) a variant has across haplotypes has a dramatic impact on the number of unique microhomology annotations (vertical axis). Transparency and jittering (± 0.5) separates points falling on integers.

Supplemental Figure 5



Supplemental Figure 5. 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.

Supplemental Figure 6



Supplemental Figure 6. 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. The correct location is difficult to identify, and all three methods chose different breakpoints.

Supplemental Figure 7

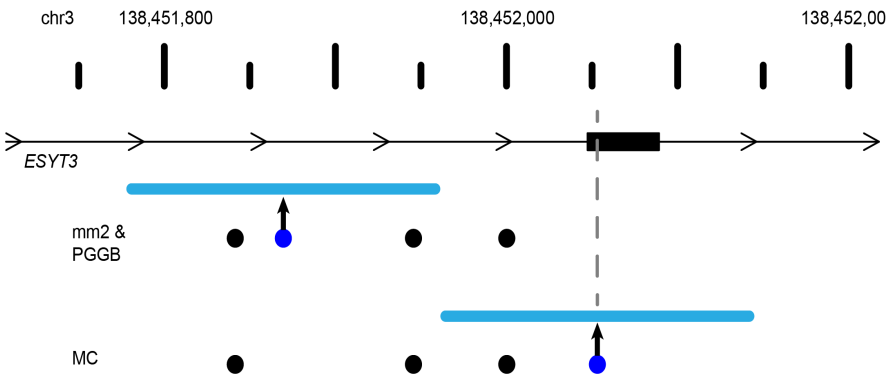


Figure S7: Breakpoint differences affect biological interpretation. A 180 bp insertion was called by minimap2 (mm2) and Pangenome Graph Builder (PGGB) at the same location within the intron of *ESYT3*, but Minigraph-Cactus (MC) placed the insertion 183 bp upstream inside an exon of *ESYT3*. SV insertions 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.

Supplemental Figure 8

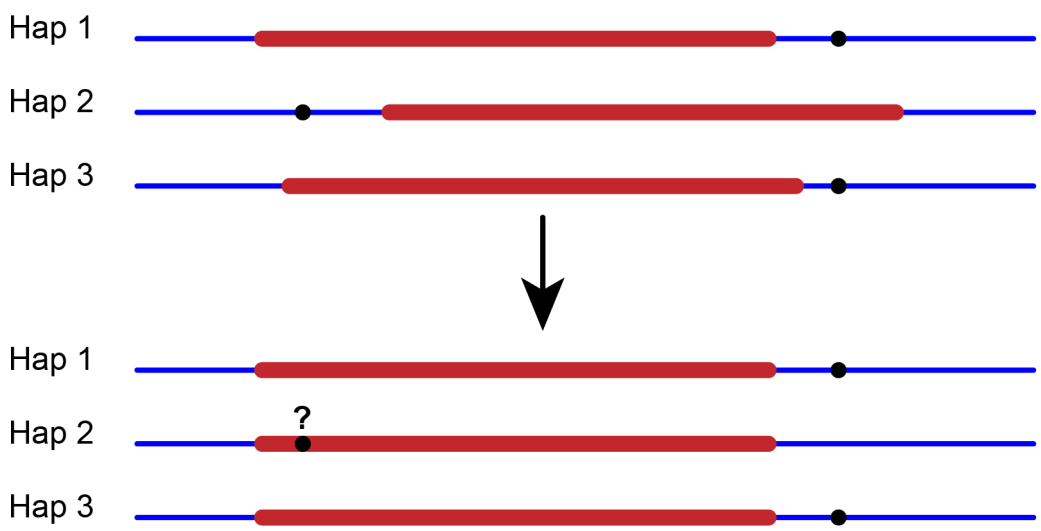


Fig S8. Over-normalizing SVs during merging. A deletion (red lines) called in three haplotypes (top three lines) is set at different breakpoints and surrounded by SNPs (black dots). When merging events into a single call, there is no way to choose a representation of the SV from one of the haplotypes without making a SNP fall inside the deletion (question mark).