Pangenome Graph Construction from Genome Alignment with Minigraph-Cactus

Glenn Hickey[1]*, Jean Monlong[1]*, Adam Novak[1], Jordan M. Eizenga[1], Human Pangenome Reference Consortium[#], Heng Li[2,3], Benedict Paten[1]

* These authors contributed equally

[1] UC Santa Cruz Genomics Institute, University of California, Santa Cruz, 1156 High St, Santa Cruz, CA, USA

[2] Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA 02215, USA

[3] Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02215, USA

[#] Full author listing appears at the end of the manuscript

Abstract

Reference genomes provide mapping targets and coordinate systems but introduce biases when samples under study diverge sufficiently from them. Pangenome references seek to address this by storing a representative set of diverse haplotypes and their alignment, usually as a graph. Alternate alleles determined by variant callers can be used to construct pangenome graphs, but thanks to advances in long-read sequencing, high-quality phased assemblies are becoming widely available. Constructing a pangenome graph directly from assemblies, as opposed to variant calls, leverages the graph's ability to consistently represent variation at different scales and reduces biases introduced by reference-based variant calls. Pangenome construction in this way is equivalent to multiple genome alignment, which is extremely computationally demanding. Here we present the Minigraph-Cactus pangenome pipeline and demonstrate its ability to build a pangenome graph by aligning 90 human haplotypes. This tool was designed to build graphs with as much information as possible while still being practical for use with current mapping and genotyping tools. We show that this graph is useful both for studying variation within the input haplotypes, but also as a basis for mapping and variant calling. We also demonstrate that this method can apply to nonhuman data by showing improved mapping and variant detection sensitivity with a *Drosophila melanogaster* pangenome.

Introduction

The term pangenome has historically referred to the set of genes present across a population or species. The patterns of presence and absence of genes from the pangenome in individual

samples, typically prokaryotes, provided a rich context for better understanding the genes and populations in question (Eizenga, Novak, Sibbesen, et al., 2020). Eukaryotic genomes can likewise be combined into pangenomes, which can be expressed in terms of genomic content rather than genes. Eukaryotic pangenomics is growing in popularity, due it part for its potential to reduce reference bias (Miga & Wang, 2021). A eukaryotic pangenome can be represented as a set of variants against a reference (Garrison et al., 2018), but technological advances in long-read sequencing are now making it possible to produce high-quality genome assemblies of samples under study, allowing for variation to be studied within its full genomic context (Abel et al., 2020). Two themes that have emerged from this work are that 1) relying on a single reference genome can be a source of bias, especially for short-read sequencing projects, and 2) representation of structural variation is a challenging problem in its own right. Pangenomes and the software toolkits that work with them aim to address these issues.

Sequence-resolved pangenomes are typically represented using graph models. There are two main classes of graph representation: sequence graphs and de-Bruijn graphs, and several different methods have been published for each type. This is an area of active research; different methods perform better for different applications, and there is as yet no clear best practice. However, sequence graphs have generally proved more amenable for read mapping (Garrison et al., 2018; Hickey et al., 2020; Sirén et al., 2021), and they will be the focus of this work. In a sequence graph, each node corresponds to a DNA sequence (**Figure 1A**) or its reverse complement depending on the direction in which it is traversed. Sample haplotypes are stored as paths, and edges are bidirected to encode strandedness (i.e. if an edge is incident to the forward or reverse complement sequence of a node). Sites of variation appear as bubbles, formally termed snarls, in the graph (Paten et al., 2018). Two snarls are indicated in the example graph in **Figure 1A,** the left and right representing a two-base substitution and 19-base deletion, respectively.

Phased Variant Call Format (VCF) files can be thought of as sequence graphs. The vg toolkit makes this perspective explicit by supporting graph construction from VCF (Garrison et al., 2018). Using such graphs for mapping and variant calling reduces reference bias and improves accuracy over GRCh38 (Garrison et al., 2018; Sirén et al., 2021). These graphs can also be used to accurately genotype structural variants (SVs) (Hickey et al., 2020), but they are still limited to reference-based variant calls. For example, there is no satisfactory way in VCF 4.3 to directly represent variation nested within a large insertion. Now that they are becoming widely available (Rautiainen et al., 2022), high-quality assemblies can be used to directly construct a pangenome graph without the need to go through variant calls. This is equivalent to finding a whole genome multiple alignment, which is known to be an extremely computationally challenging problem (Just, 2004). As such, multiple alignment algorithms must use heuristics for scaling with respect to both the number of input sequences and their combined length. Typically, the former is accomplished by decomposing the multiple alignment of N genomes into N pairwise alignments and the latter by seed-and-extend heuristics (Kille et al., 2022).

MultiZ was among the first methods able to align dozens of vertebrate genomes and is still used by the UCSC Genome Browser. It begins with a set of pairwise alignments of the input genomes

to a given reference assembly, then uses progressive decomposition to merge the alignments according to their phylogenetic relationships (Blanchette et al., 2004). The pairwise alignments themselves are created with LASTZ, which uses a gapped seeding approach to find anchors, which are then chained and extended with dynamic programming (Harris, 2007). Progressive Cactus is a more recent tool for large vertebrate scale multiple alignments (Armstrong et al., 2020). It also uses LASTZ, or the GPU-accelerated successor SegAlign (Goenka et al., 2020), to perform pairwise alignments. However, it does so by progressively reconstructing ancestral sequences using a phylogenetic guide tree. This eliminates the need for a global reference assembly, making Progressive Cactus reference-independent. At each step, the LASTZ alignments are used as anchors to construct a Cactus graph (Paten, Diekhans, et al., 2011), which in turn is used to filter and then refine the alignment. Progressive Cactus was used to produce an alignment of 600 amniotes: still, to our knowledge, the largest vertebrate alignment ever computed.

Progressive Cactus was shown to be robust to small errors in the guide tree, but, like any progressive alignment approach, it still relies upon an accurate phylogenetic tree. Due to recombination, a single tree cannot reasonably represent the the ancestry of any intraspecies genome set that one might want to use to construct a pangenome. Minigraph (Li et al., 2020) is a newer tool that uses an iterative sequence-to-graph mapping approach, similar to Partial Order Alignment (POA) (Lee et al., 2002), to construct a pangenome graph from a set of input genomes. It uses a generalization of minimap2's minimizer-based seeding and chaining strategy (Li, 2018), and is similarly fast so long as the input genomes are relatively similar. While minigraph can perform base-level alignment since version 0.17, it only includes SVs (≥ 50bp by default) during graph construction. Excluding small variation prevents input genomes from being losslessly embedded as paths in the graph.

We now present the Minigraph-Cactus pangenome pipeline, which combines Minigraph's fast assembly-to-graph mapping with Cactus's base aligner in order to produce base-level pangenome graphs at the scale of hundreds of vertebrate haplotypes. In addition to representing variation consistently at all resolutions, we show that these graphs can be used to improve short-read mapping, variant calling, and SV genotyping.

The Minigraph-Cactus Pangenome Pipeline

The Minigraph-Cactus Pangenome pipeline is included in Cactus. Like Cactus, it is implemented using Toil (Vivian et al., 2017), which allows it to be run both locally and distributed in the cloud. The pipeline consists of five steps as shown in **Figure 1B**, which are used to generate a graph in both GFA and VCF format, as well as indexes required to map reads using vg giraffe (Sirén et al., 2021).

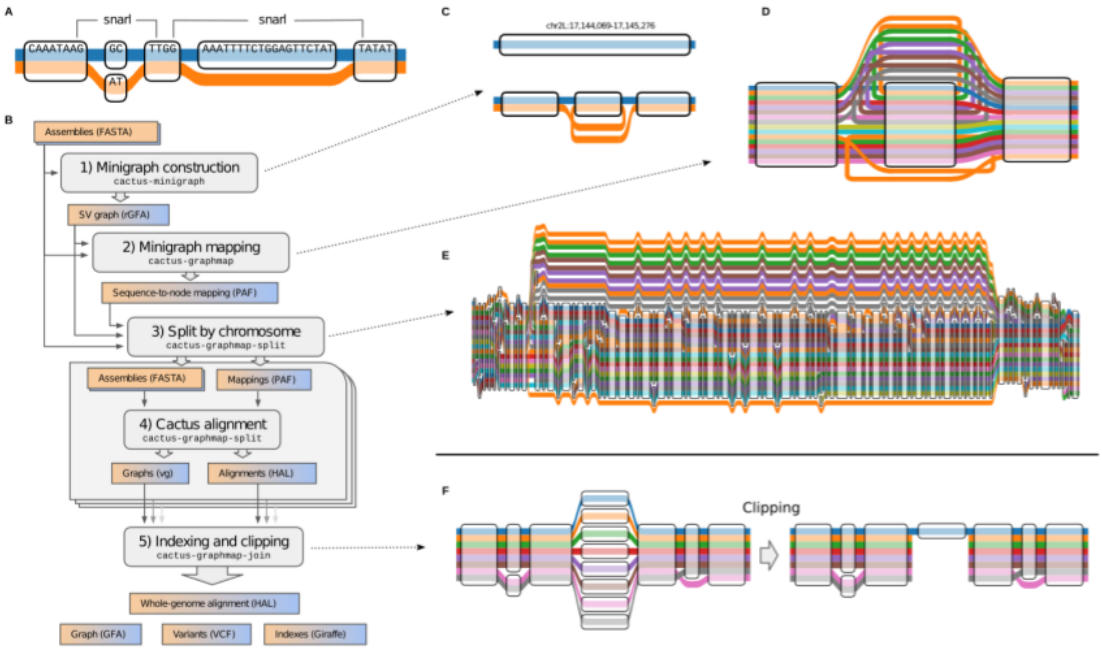


Figure 1: A) “Tube Map" view of a sequence graph shows two haplotypes as paths through the graph. The two snarls (variation sites defined by graph topology, aka bubbles) are highlighted. **B)** The five steps, and associated tools, of the Minigraph-Cactus pipeline which takes as input genome assemblies in FASTA format and outputs a pangenome graph, genome alignment, VCF and indexes required for mapping with vg Giraffe. Illustrating the steps in the pipeline by example: **C)** SV graph construction using minigraph (as wrapped by cactus-minigraph) begins with a linear reference and adds SVs, in this case a single 1204bp inversion (at ch2L:17,144,069 in the *D. melanogaster* pangenome). **D)** The input haplotypes are mapped back to the graph with minigraph, in this example six of which contain the inversion allele from C. **E)** The minigraph mappings are combined into a base-resolution graph using Cactus, augmenting the larger SVs with smaller variants - in this case, adding smaller variants within the inversion. **F)** An unaligned centromere is clipped out of a graph, leaving only the reference (blue) allele in that region. The other alleles are each broken into two separate subpaths but are otherwise unaffected outside the clipped region.

Minigraph SV graph construction

The pipeline begins with the construction of an initial SV-only graph using `minigraph -xggs -c`, as described in (Li et al., 2020). This is an iterative procedure that closely resembles partial order alignment (POA): a “reference” assembly is chosen as an initial backbone, and then augmented with variation from the remaining assemblies in turn. **Figure 1C** shows an example of an inversion being augmented into a reference chromosome. Minigraph does not collapse duplications: If two copies of a gene are present in the graph after adding *i* genomes, but there are three copies in the *i+1*st genome, then an additional copy will be added to the graph. This is

a key difference between minigraph and other approaches (including Progressive Cactus) that would tend to collapse all copies of the gene into a single sequence in the absence of outgroup information to determine the ancestral state. By keeping different gene copies separate, minigraph trades greater graph complexity (more bubbles) for reduced path complexity (fewer cycles).

Minigraph contig mapping

Minigraph generalizes the seeding and chaining concepts from minimap2 (Li, 2018) for use on sequence graphs and, since version 0.17, can produce base-level alignments between contigs and graphs (but not base-level *graphs*). In this step of the pipeline each assembly, including the reference, is mapped back to the SV graph independently using minigraph's `-xasm -c` options (**Figure 1D)**. The results are concatenated into a single GAF file, which is then filtered to remove spurious alignments. By re-aligning each assembly to the same graph in this step as opposed to re-using mapping created during construction, we mitigate an issue in the latter where orthologous sequences can be aligned to inconsistent locations when mapped to different versions of the graph. We remove any aligned query interval that overlaps another by at least 25% of its length, and whose mapping quality and/or block length is 5X lower than those of the other interval. For human data, an alignment is also filtered out if it would induce a novel deletion of ≥ 10Mb in the graph. This prevents spurious deletion edges, potentially caused by misassembly or misalignment. Finally, the GAF is converted into PAF format by splitting up records so that each line has, for a target, a single node sequence.

Splitting by chromosome

Minigraph does not introduce interchromosomal events, so every node in the graph is connected to exactly one reference component. This information, obtained via the rGFA "SN" tag, allows the graph to be decomposed into a connected component for each reference scaffold (or chromosome if the reference is assembled into chromosomes). The contig alignments can be decomposed similarly. Contigs that map to multiple components are assigned to the one with the most coverage, while contigs that do not sufficiently map anywhere are considered ambiguous and left out of further analysis. Splitting by chromosome ensures that Cactus will not introduce interchromosomal events, many of which would be alignment artifacts for human data. It also limits memory usage in Cactus, but is only strictly necessary to split by chromosome when resource limits require which, roughly speaking, would be after about 30Gb of input sequence for 256 Gb of RAM. Contigs are only assigned to a chromosome if at least a certain fraction of their bases align to that chromosome, and this fraction is at least three times greater than that of any other chromosome. The fraction is 75% for contigs with length ≤ 100 kb, 50% for contigs with length in the range (100 kb, 1 Mb] and 25% for with length > 1 Mb. Any contig that does not meet these criteria is flagged as "ambiguous" and excluded from the graph.

Cactus base alignment

Cactus combines a set of pairwise alignments into a multiple alignment (Armstrong et al., 2020; Paten, Earl, et al., 2011). This process remains conceptually unchanged whether we input the Minigraph mappings or the all-to-all LASTZ mappings used in the standard Cactus pipeline: It begins by "pinching" exactly matching aligned bases together to form an initial sequence graph, where every node is a sequence fragment and can be considered to have two sides (**Figure 1A**). This sequence graph is then transformed into a Cactus graph, which represents the chains of alignment within the sequence graph (Paten, Diekhans, et al., 2011). The topology of the Cactus graph is first used to remove candidate spurious or incomplete alignments corresponding to short, high-degree alignment chains. Interstitial unaligned sequences that share common anchors are then aligned together. Cactus originally used the cPecan multiple sequence aligner for this process (Paten et al., 2008), but that has now been changed to abPOA (Gao et al., 2021), primarily for computational reasons. abPOA is more than twice as fast for small alignments, and it scales linearly with the number of sequences being aligned, whereas cPecan scales quadratically.

Cactus natively outputs genome alignments in Hierarchical Alignment (HAL) format (Hickey et al., 2013). HAL files can be used to create assembly hubs on the UCSC genome browser, or to map annotations between genomes (Fiddes et al., 2018), but they are not suitable for most pangenome graph applications, which expect GFA. An option was therefore added to Cactus to additionally output the alignment as a sequence graph in vg and/or GFA format. These graphs contain the underlying structural variation from the minigraph along with smaller variants, and the input haplotypes are represented as paths (**Figure 1E**).

Indexing and clipping

The final step of the pipeline combines the chromosome level results and performs some post-processing. This includes resolving conflicting node ids using `vg ids -j` and collapsing redundant nodes where possible using `gaffix` (Doerr, 2022). Nodes are also replaced with their reverse complement as necessary to ensure that reference paths only ever visit them in the forward orientation. The original SV graph produced by minigraph remains embedded in the results at this stage, with each minigraph node being represented by a separate embedded path.

Minigraph-Cactus cannot presently map through highly repetitive sequences like satellite arrays, centromeres and telomeres. As such, these regions will remain largely unaligned throughout the pipeline and will make the graph difficult to index and map to by introducing vast amounts of redundant sequence. We recommend clipping them out for most applications and provide the option to do so by removing paths with >N bases that do not align to the underlying minigraph (**Figure 1F**). In preliminary studies of mapping short reads and calling small variants (see below), we found that even more aggressively filtering the graph helps improve accuracy. For

this reason, an optional allele-frequency filter is included to remove nodes of the graph present in fewer than N haplotypes.

In all, up to three graphs are produced:

1) Full graph: useful for storing complete sequences and performing liftover; difficult to index and map to because of unaligned centromeres
2) Default graph: clip out all stretches of sequences >=10kb that do not align to the minigraph. The intuition is that large SVs not in minigraph are under-alignments of sequence not presently alignable and not true variants. This graph is ideal for studying variation and exporting to VCF, and can be effectively indexed for read mapping.
3) Allele-frequency filtered graph: remove all nodes present in fewer than N haplotypes. This filter increases accuracy for short read mapping and variant calling, as shown in **Supplementary Figures 3** and **4**, respectively.

Graph 2) is a subgraph of graph 1), and graph 3) is a subgraph of graph 2). They are node-id compatible, in that any node shared between two of the graphs will have the same sequence and ID. Unless otherwise stated, all results below about the graphs themselves are referring to the default graphs, whereas all results pertaining to read mapping and genotyping were performed on the allele-frequency filtered graphs.

Human Pangenome Reference Graphs

The Minigraph-Cactus pipeline was originally developed to construct a pangenome graph for the assemblies produced by the Human Pangenome Reference Consortium (HPRC). In its first year, this consortium released 47 diploid assemblies (Liao et al., 2022). For evaluation purposes, we held out three samples when generating the graph: HG002, HG005 and NA19240. The remaining 44 samples (88 haplotypes), and two reference genomes (GRCh38 and CHM13 [v1.1] (Nurk et al., 2022) were used to construct the graph, with 90 haploid genomes total. Since the construction procedure is dependent on the reference chosen for the graph, we ran our pipeline twice independently on the same input assemblies, once using GRCh38 as the reference and once CHM13. The CHM13-based graph includes more difficult and highly variant regions, such as in the acrocentric short arm of chr21, that are not represented in the GRCh38-based graph. This makes it slightly bigger than the GRCh38-based graph, both in terms of total sequence and in terms of nodes and edges (**Supplementary Table 1**). The final pangenomes have roughly 200X more nodes and edges than the SV Graphs from Minigraph, showing the amount of small variation required in order to embed the haplotype paths. **Figure 2A** shows the amount of non-reference sequence as a function of how many haploid genomes contain it (the same plot for total sequence can be found in **Supplementary Figure 1**). The rise in the leftmost points (support=1) is due to private sequence, only present in one sample, and may also contain alignment artifacts which often manifest as under-alignments affecting a single sample. The plot clearly shows that the CHM13-based graph has less non-reference sequence present across the majority of samples, an apparent consequence of the improved completeness of CHM13 over GRCh38. The distribution of allele sizes within

snarls (variant sites in the pangenome defined by graph topology; **Figure 2B**) highlights the amount of small variation added relative to Minigraph alone. The total time to create and index each HPRC pangenome graph was roughly 3 days (**Supplementary Table 3**).

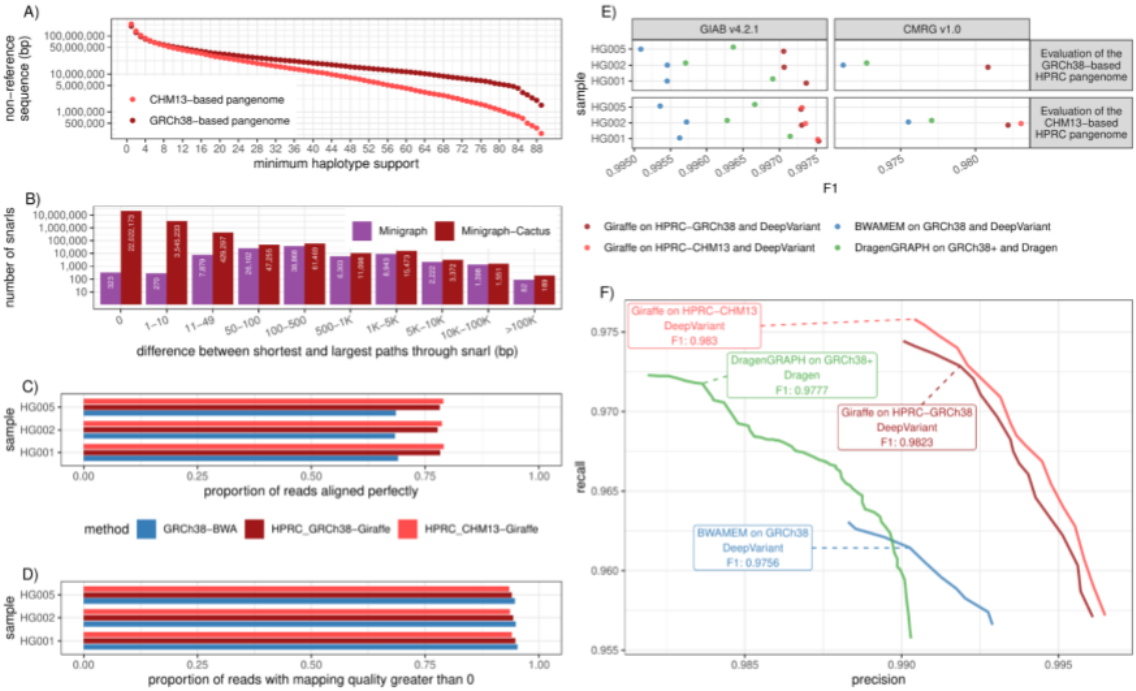


Figure 2. A) The amount of non-reference sequence in the HPRC graphs by the minimum number of haplotypes it is contained in. **B)** Distribution of the size of the snarls (variation sites) for the GRCh38-based Minigraph and Minigraph-Cactus pangenomes. **C-D)** ~30x Illumina short-reads for three GIAB samples were mapped using three approaches: BWA-MEM on GRCh38 (red), vg Giraffe on the GRCh38-referenced HPRC pangenome (blue) and vg Giraffe on the CHM13-referenced pangenome (green). **C)** Proportion of the reads aligning perfectly to the (pan-)genome for each sample (y-axis). **D)** Proportion of reads with a mapping quality greater than 0. **E-F)** Short variants were called with DeepVariant after projecting the reads to GCRh38 from the GRCh38-based pangenome (dark red), or the CHM13-based pangenome(light red). The results when aligning reads with BWA-MEM (blue) or using the Dragen pipeline (green) are also shown. E) The F1 score is shown on the x-axis across samples from the Genome in a Bottle (y-axis). Left: Genome in a Bottle v4.2.2 high confidence calls. Right: Challenging Medically Relevant Genes v1.0. When evaluating the CHM13-based pangenome (bottom panels), regions with false duplications or collapsed in GRCh38 were excluded. F) The graph shows the precision (x-axis) and recall (y-axis) for different approaches using the Challenging Medically Relevant Genes v1.0 truth set for the HG002 sample (bottom-right panel in E)). The curves are traced by increasing the minimum quality of the calls.

Mapping to the HPRC Graphs

We benchmarked how well the pangenome graphs could be used as drop-in replacements for linear references in a state-of-the-art small variant discovery and genotyping pipeline. To do so, we used Illumina short read sequence data with about 30x coverage from three Genome in a Bottle (GIAB) samples, HG001, HG002, and HG005. All mapping experiments were performed on filtered HPRC graphs with a minimum allele frequency of 10%, meaning that nodes supported by fewer than 10 haplotypes were removed. We found that reads aligned with higher identity when mapped to the pangenomes using Giraffe, compared to the traditional approach of mapping reads with BWA-MEM on GRCh38. On average, 78.1% and 78.9% of reads aligned perfectly for the GRCh38-based and CHM13-based pangenomes, respectively, compared to 68.7% when using BWA-MEM on GRCh38 (**Figure 2C**). Mapping to the pangenomes results in a slight drop in mapping confidence, from about 94.9% to 94.1% of reads with a mapping quality greater than 0 (**Figure 2D**) in those samples. This is expected as the pangenome contains more sequence than GRCh38, including complex regions and large duplications that are now more fully represented, which naturally and correctly reduces mapping confidence for some reads. The same trend is observed when the pangenome is not filtered by frequency, although slightly fewer reads map with a mapping quality greater than 0 (**Supplementary Figure 4**).

Variant Calling with the HPRC Graphs

We used the short read alignments to call variants with DeepVariant (Poplin et al., 2018). To prepare them for DeepVariant, the graph alignments were projected onto GRCh38 using the vg toolkit. Note that, even though the CHM13-based graph did not use GRCh38 as the initial reference, the graph does contain GRCh38. Thus, the CHM13-based graph can also be used in this pipeline.

Both pangenomes constructed with Minigraph-Cactus outperform the other methods (**Figure 2E-F**). We note that reads in regions that are falsely duplicated or collapsed in GRCh38 cannot be unambiguously projected from their corrected alleles in CHM13. For this reason, these regions were removed from the benchmark when evaluating the CHM13-based pangenome. Unsurprisingly, the CHM13-based pangenome offers the largest gains in variant calling in challenging regions like those assessed by the Challenging Medically Relevant Genes truth set (**Figure 2F**) (Wagner et al., 2022). Of note, the frequency-filtered pangenomes performed better than using the default pangenomes (**Supplementary Figure 5**). We also tested projecting and calling variants on CHM13. Although the benchmarking protocol is still preliminary for CHM13, we observed a clear improvement when using the pangenome compared to aligning the reads to CHM13 only (**Supplementary Figure 6**). Some specific regions, including the MHC region and segmental duplications, seem to have better variant calls on the CHM13-based graph (**Supplementary Figure 7**).

D. Melanogaster Pangenome

We created a *Drosophila melanogaster* pangenome to demonstrate Minigraph-Cactus's applicability to non-human organisms. We used 16 assemblies including the reference, dm6 (ISO1), 14 geographically diverse strains described in (Chakraborty et al., 2019), and one additional strain, B7. Their sizes range from 132 to 144 Mb. The allele frequency filtered graph, used for all mapping experiments, was created by removing nodes appearing in < 2 haplotypes leading to a minimum allele frequency of ~12.5% (compared to 10% in the human graph), and was used only for mapping and genotyping, where private variation in the graph is less helpful. The relatively small input meant that we could align it with Progressive Cactus using an all-vs-all (star phylogeny) rather than progressive alignment, and the results are included for comparison. In all, we produced five *D. melanogaster* graphs whose statistics are shown in **Supplementary Table 2**, a process that took roughly 5 hours (**Supplementary Table 3)**. As in human, adding base-level variants to the SV graph increases its number of nodes and edges by roughly two orders of magnitude. Progressive Cactus, which is not tuned for such large all-to-all alignments, has far more nodes and edges, and less aligned with a greater total node length. It does, however, contain all input contigs and allows interchromosomal events.

The "core" genome size, which we define as the total length of all nodes present in all samples, of the Minigraph-Cactus pangenome is 110 Mb (**Supplementary Figure 8**, first column), which is roughly half the total size of the graph. This reflects a high diversity among the samples: private transposable element (TE) insertions are known to be abundant in this species (Chakraborty et al., 2019). This diversity is also shown in **Figure 3A**, which graphs the amount of non-reference sequence by the minimum number of samples it is present in, where the private TE insertions would account for much of the nearly 10X differene between the first and second columns. The trend for the number of non-reference nodes is less pronounced (**Supplementary Figure 9**), which implies that the non-reference sequence is accounted for by larger insertion events and smaller variants tend to be more shared. We used the snarl subgraph decomposition (Paten et al., 2018) to compute the variant sites within each graph, i.e. subgraphs equivalent to individual SNPs, indels, SVs, etc. **Supplementary Figure 10** shows the pattern of nesting of the variant sites in the various graphs.

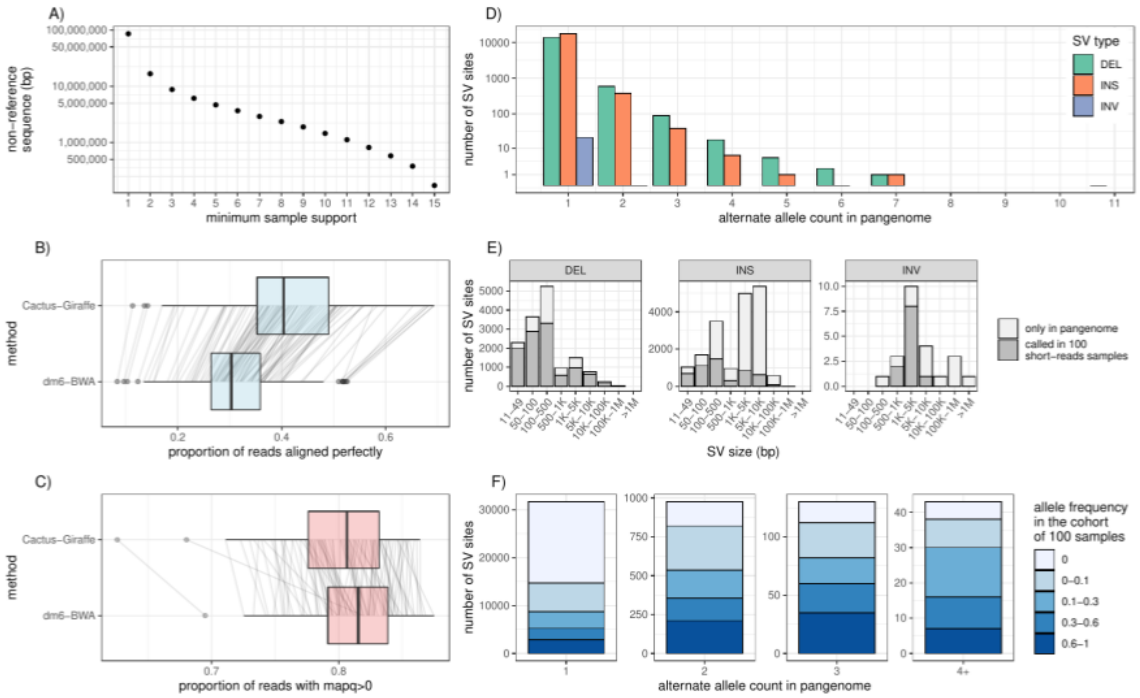


Figure 3: A) Amount of non-reference sequence by minimum number of haplotypes it occurs in for the *D. melanogaster* pangenome. **B)** Proportion of reads that align perfectly (x-axis) to the filtered pangenome for two approaches (x-axis): "Cactus-Giraffe" where short reads are aligned to the pangenome using vg Giraffe; "dm6-BWA" where reads were mapped to dm6 using BWA-MEM. Each point represents a sample. The lines connect a same sample between the two approaches. **C)** Proportion of reads with a mapping quality above 0. **D)** Distribution of the alternate allele count across each SV site. The x-axis represents the number of assemblies in the pangenome that support a SV. The y-axis is log-scaled. **E)** The size distribution (x-axis) of different SV types (panels). The SV sites are separated in two groups: SV sites that were called in at least one sample from the cohort of 100 samples with short reads (dark grey); SV sites only present in the pangenome (light grey). **F)** Fraction of SVs of different frequency in the cohort of 100 samples (color) compared to their frequency in the pangenome (x-axis).

Short-read Mapping

The *Drosophila melanogaster* Genetic Reference Panel (DGRP) consists of 205 inbred genomes (Huang et al., 2014), unrelated to the 16 strains used to construct the pangenome. We used short reads from this dataset to evaluate mapping performance for our pangenome graph. We selected 100 samples for our evaluation, filtering the dataset to include only samples with a single SRA accession and Illumina sequencing with >15X coverage. We mapped these samples to the allele frequency filtered pangenome graph with vg giraffe in "fast" mode, and to dm6 using BWA-MEM. We counted the number of mapped reads, reads with perfect alignment, and reads with a mapping quality above 0. We found that the number of reads aligning perfectly drastically increased (**Figure 3B**), with on average 41.1% of the reads aligning perfectly to the pangenome compared to on average 31.0% when aligning reads with BWA on dm6. As in our results in human presented above, we observe a decrease in the number of reads mapped with

a mapping quality above 0 when mapping to the pangenome (80.0% vs 81.1% on average, **Figure 3C**).

Small Variants

We used projected pangenomic mappings to dm6, and used FreeBayes (Garrison & Marth, 2012) (in the absence of a high quality DeepVariant model) to call variants on these mappings and those from BWA-MEM (see Methods). We then compared the variant calls that were called by both approaches, and those that were called by only one. While variant sites called by both methods showed similar quality scores, there were more sites unique to our pangenomic approach compared to sites found only by mapping reads to the linear dm6 genome. This increase was observed across different quality thresholds (**Supplementary Figure 12A,C**). Overall, that meant that slightly more variants are called when mapping short reads to the pangenome and projecting them to dm6. For example, on average 740,696 small variants had a quality above 0.1 compared to 738,570 when reads were mapped to the dm6 with BWA-MEM (**Supplementary Figure 12B**). For genotype quality above 10, 705,320 small variants were called versus 700,385 (**Supplementary Figure 12D**). We also noticed a lower rate of heterozygous variants called when mapping the reads to the pangenome first (13.2% vs 18.1% on average per sample, **Supplementary Figure 13**). Due to the high inbreeding of these samples, we expect only a small fraction of variants to truly be segregating (Huang et al., 2014).

Structural Variants

The variant sites in the pangenome (snarls) were decomposed into canonical structural variants based on the assembly paths in the pangenome (see Methods). In the pangenome, most of the SVs are rare and supported by one or two assemblies (**Figure 3D**). Of note, the known In(3R)C inversion (Miller et al., 2020) is present in the pangenome, along with 23 other smaller inversions. Structural variants were also genotyped from the short read alignments to the pangenome using vg (Hickey et al., 2020) (see Methods). Even though the genotyping used short reads and the pangenome was frequency-filtered, 47.8% of the SVs in the pangenome were found when genotyping the 100 samples (on the filtered pangenome) with short-read data. Both the full set of SVs in the pangenome and the subset genotyped from the short read data span the full size spectrum of deletions, insertions and a few inversions (**Figure 3E**). As expected, SVs that were seen in multiple assemblies in the pangenome tended to have higher allele frequencies in the cohort of 100 samples (**Figure 3F**). Both rare and more common SVs spanned the full spectrum of SV size and repeat profile, from the shorter simple repeats and satellite variation to the larger transposable element polymorphisms of LTR/Gypsy, LTR/Pao, and LINE/I-Jockey elements, among others (**Supplementary Figure 14**).

Discussion

It will soon be routine to produce large numbers of reference quality genomes for many species. The HPRC has shown that, for human, such data can be leveraged to provide a better window into structural variation, as well as to reduce bias incurred by relying on a single reference. The pangenome graph representation has been fundamental to this work, but graph construction remains an active research area. The key challenges stem not just from the computational difficulty of multiple genome alignment, particularly in complex regions, but also from fundamental questions about the tradeoffs between complexity and usability. While developing Minigraph-Cactus, we sought a method to construct graphs with as much variation as possible, while still serving as useful inputs for current pangenome tools like vg and PanGenie (Ebler et al., 2022). Graph construction, including all indexing, took under 3 days for the HPRC graphs using an AWS cluster, and 5 hours for *D. Melanogaster* graph on a single computer.

Leveraging the ability to project read mappings to different references, we also showed how pangenomes can be used to improve variant calling on linear references, which we expect to be key to fostering adoption of pangenomes in the near term. Similarly important, we think, is the ability to handle multiple references, which we demonstrated by including CHM13 and GRCh38 in the HPRC graphs. Minigraph-Cactus allows accurate variant calling pipelines on both. Some of the compromises made to make our method practical represent exciting challenges for future work in both pangenome construction and applications. Pangenomes from Minigraph-Cactus cannot be used, for instance, to study centromeres. The omission of interchromosomal events will likewise preclude useful cancer pangenomes or studies into acrocentric chromosome evolution (Guarracino et al., 2022). We are also interested in ways to remove the necessity of filtering the graph to get optimal mapping performance by using an online method at mapping time to identify a subgraph that most closely relates to the reads of a given sample.

Pangenomics has its origin in non-human organisms, and as the assembly data becomes available, we will see pangenomes being produced for a wide array of organisms. Already there is data for a number of species, from tomato (Zhou et al., 2022) to cow (Leonard et al., 2022). In this work, we constructed a *D. melanogaster* pangenome as a proof of concept to show that our method can also be used on other non-human organisms. We hope that others will use the Minigraph-Cactus pipeline to produce useful graphs from sets of genome assemblies for their species of interest. And, as we expect pangenome references to supplant single genome references for intraspecies population genomics studies, we also see this as the future in interspecies comparative genomics studies.

Methods

Software Availability

Minigraph-Cactus is included in Cactus, which is released as source, static binaries and Docker images here: https://github.com/ComparativeGenomicsToolkit/cactus/releases. The user guide is here: https://github.com/ComparativeGenomicsToolkit/cactus/blob/mc-paper-sub/doc/pangenome.md. Links to the human and *D. melanogaster* pangenome graphs and indexes, as well as those for some other species can be found here: https://github.com/ComparativeGenomicsToolkit/cactus/tree/mc-paper-sub/doc/mc-pangenomes/README.md. Please consult https://github.com/ComparativeGenomicsToolkit/cactus/blob/mc-paper-sub/doc/mc-paper/README.md for command lines and scripts used for this work.

HPRC Graph Construction

The HPRC v1.0 graphs discussed here were created by an older version of the pipeline described above, with the main difference being that the satellite sequence was first removed from the input with dna-brnn (Li, 2019). This procedure is described in detail in (Liao et al., 2022). The amount of sequence removed from the graph, and the reason it was removed, is shown in **Supplementary Figure 2**. Roughly 200 Mb per assembly was excluded, the majority of which was flagged as centromeric (HSat2 or alpha satellite) by dna-brnn (Li, 2019). The “unassigned”, “minigraph-gap” and “clipped” categories denote the sequence that, respectively, did not map well enough to any one chromosome to be assigned to it, intervals > 100kb that did not map with minigraph, and intervals > 10kb that did not align with Cactus. Simply removing all sequence ≥10kb that does not align with Cactus, as described in the methods above, amounts to nearly the same amount of sequence excluded (**Supplementary Figure 3**). The exact commands to build HPRC graphs referred to in this figure are available here: https://github.com/ComparativeGenomicsToolkit/cactus/blob/91bdd83728c8cdef8c34243f0a52b28d85711bcf/doc/pangenome.md#hprc-graph. They were run using the same Cactus commit: 91bdd83728c8cdef8c34243f0a52b28d85711bcf.

By default, all graphs are output in GFA (v1.1), as well as the vg-native indexes: xg, snarls and gbwt formats (Eizenga, Novak, Kobayashi, et al., 2020; Sirén et al., 2020). Reference-based variants in VCF format can also be produced. This is accomplished by using `vg deconstruct`, which outputs a site for every snarl in the graph. It uses the haplotype index (GBWT) to enumerate all haplotypes that traverse the site, which allows it to compute phased genotypes. For each allele, the corresponding path through the graph is stored in the AT (Allele Traversal) tag. Snarls can be nested, and this information is specified in the LV (Level) and PS (Parent Snarl) tags, which needs to be taken into account when interpreting the VCF.

HPRC Graph Mapping and Variant Calling

We used 30x Illumina NovaSeq PCR-free short read data HG001, HG002, and HG005, available at gs://deepvariant/benchmarking/fastq/wgs_pcr_free/30x/. The reads were mapped to the pangenome using vg giraffe (v1.37.0). The same reads were mapped to GRCh38 with decoy sequences, but no ALTs using BWA-MEM (v0.7.17). The number of reads mapped with

different mapping quality (or aligning perfectly) were extracted from the graph alignment file (GAF/GAM files) produced by vg giraffe and from the BAM files produced by BWA-MEM.

Variants were called using the approach described in (Liao et al., 2022). Briefly, the graph alignments were projected to the chromosomal paths (chr 1-22, X, Y) of GRCh38 using vg surject. Once sorted with samtools (v1.3.1), the reads were realigned using bamleftalign (Freebayes v1.2.0) (Garrison & Marth, 2012) and ABRA (v2.23) (Mose et al., 2014). DeepVariant (v1.3) (Poplin et al., 2018) then called small variants using models trained for the HPRC pangenome (Liao et al., 2022). We used the same approach when calling small variants using the CHM13-based pangenome and when projecting to CHM13 chromosomal paths.

Evaluation of small variant calls

Calls on GRCh38 were evaluated as in (Liao et al., 2022), i.e. using the Genome In A Bottle (GIAB) benchmark and confident regions for each of the three samples (Zook et al., 2016). For HG002, the Challenging Medically Relevant Genes (CMRG) truth set v1.0 (Wagner et al., 2022) was also used to evaluate small variants calls in those challenging regions. The evaluation was performed by hap.py (Krusche et al., 2019) v0.3.12 via the jmcdani20/hap.py:v0.3.12 docker image.

When evaluating calls made against the GRCh38 chromosomal paths using the CHM13-based pangenome, we excluded regions annotated as false-duplications and collapsed in GRCh38. These regions do not have a well-defined truth label in the context of CHM13. We used the “GRCh38_collapsed_duplication_FP_regions”, “GRCh38_false_duplications_correct_copy”, “GRCh38_false_duplications_incorrect_copy”, and “GRCh38_population_CNV_FP_regions” region sets available at https://github.com/genome-in-a-bottle/genome-stratifications.

To evaluate the calls made on CHM13 v1.1, we used two approaches. First, the calls from CHM13 v1.1 were lifted to GRCh38 and evaluated using the GRCh38 truth sets described above (GIAB v4.2.1 and CMRG v1.0). For this evaluation, we also lifted these GRCh38-based truth sets to CHM13 v1.1 to identify which variants of the truth set are not visible on CHM13 because they are homozygous for the CHM13 reference allele. Indeed, being homozygous for the reference allele, those calls will not be present in the VCF because there are no alternate alleles to find. These variants were excluded from the truth set during evaluation. The second approach was to evaluate the calls in CHM13 v1.1 directly. To be able to use the CMRG v1.0 truth set provided by the GIAB, we lifted the variants and confident regions from CHM13 v1.0 to CHM13 v1.1. The CMRG v1.0 truth set focuses on challenging regions, but still provides variant calls across the whole genome. Hence, we used those variants to evaluate the performance genome-wide although restricting to a set of confident regions constructed by intersecting the confident regions for HG002 from GIAB v4.2.1 (lifted from GRCh38 to CHM13 v1.1), and the alignment regions produced by dipcall in the making of the CMRG v1.0 truth set (https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/AshkenazimTrio/HG002_NA24385_son/CMRG_v1.00/CHM13v1.0/SupplementaryFiles/HG002v11-align2-CHM13v1.0/HG002v11-align2-CHM13v1.0.dip.bed). Finally, we used the preliminary HG002 truth set from GIAB on CHM13 v2.0 which is equivalent to CHM13 v1.1 with the added chromosome Y from HG002.

The calls in this set wer based on aligning a high-confidence assembly using dipcall (Li et al., 2018) (labeled in figure as “dipcall CHM13 v2.0”). Here again, we intersected the confident regions with the GIAB v4.2.1 confident regions lifted from GRCh38 to CHM13.

In all experiments described above, the variants (VCF files) were lifted over using Picard (v2.27.4) (*Broadinstitute/picard*, n.d.) LiftoverVcf and the RECOVER_SWAPPED_REF_ALT option. Regions (BED files) were lifted with liftOver (Kuhn et al., 2012).

Finally, we compared in greater detail the calling performance using the GRCh38-based and CHM13-based pangenomes by stratifying the evaluation across genomic region sets provided by the GIAB (https://github.com/genome-in-a-bottle/genome-stratifications). These regions included, for example, different types of challenging regions like segmental duplications, simple repeats, transposable elements.

D. Melanogaster Graph Construction

The *D. Melanogaster* pangenome was created using Minigraph-Cactus using the procedure described in The Minigraph-Cactus Pangenome Pipeline secion . Progressive Cactus was run on the same input (which implies a star phylogeny) and was exported to vg with `hal2vg`.

D. Melanogaster Variant Decomposition

The variant sites in the pangenome (snarls) were decomposed into canonical structural variants using a script developed for the HPRC analysis (Liao et al., 2022). In brief, each allele in the deconstructed VCF specifies the corresponding path in the pangenome. The script follows these paths and, comparing them with the dm6 reference path, enumerates each canonical variant (SNP, indels, structural variants). The frequency of each variant in the pangenome corresponds to the number of assemblies that traverse their paths.

D. Melanogaster Graph Mapping and Variant Calling

The DGPR samples used are listed in **Supplementary Table 5**. Short reads were obtained using `fasterq-dump –split 3` on the accessions in the last column of this table. Each read pair was mapped to the allele-frequency filtered graph with `vg giraffe` and to dm6 with BWA-MEM.

`vg call` was used to to genotype variants in the pangenome. For each sample, these variant calls were decomposed into canonical SVs using the same approach described above on the HPRC deconstructed VCF. The SV calls were then compared to the SVs in the pangenome using the sveval package (Hickey et al., 2020) which matches SVs based on their types, sizes and location. Here, two SVs were matched if: their regions had a reciprocal overlap of at least 90% for deletions and inversions; they were located at less than 100bp from each other, and their inserted sequences were at least 90% similar for insertions. The same approach was used

to cluster the SVs alleles into the SV sites reported in the text and figures. The SV alleles were annotated with RepeatMasker (v4.0.9). We assigned a repeat class to a SV if more than 80% of the allelic sequence was annotated as such.

We used `vg surject` to produce BAM files referenced on dm6 from the mappings to the pangenome, and FreeBayes v1.3.6 (Garrison & Marth, 2012) (in the absence of a high quality DeepVariant model) to call variants on these mappings and those from BWA-MEM. Single-sample VCFs were merged with `bcftools merge`.

To compare the variant calls by both approaches, we used bcftools (Danecek et al., 2021) (v1.10.2) to normalize the VCFs (bcftools norm), and compare them (bcftools isec) to mark variant sites where both approaches call a variant, and sites where only one approach does. We compared the number of calls in each category, across samples, and for different minimum variant quality thresholds (QUAL field or genotype quality GQ field).

Acknowledgements

We thank Yan Gao for supporting abPOA. We also thank Anthony D. Long for many suggestions and insights regarding the *D. Melanogaster* data. We thank the whole vg team for their work to create and maintain vg upon which much of this work depends.

Human Pangenome Reference Consortium Authorship

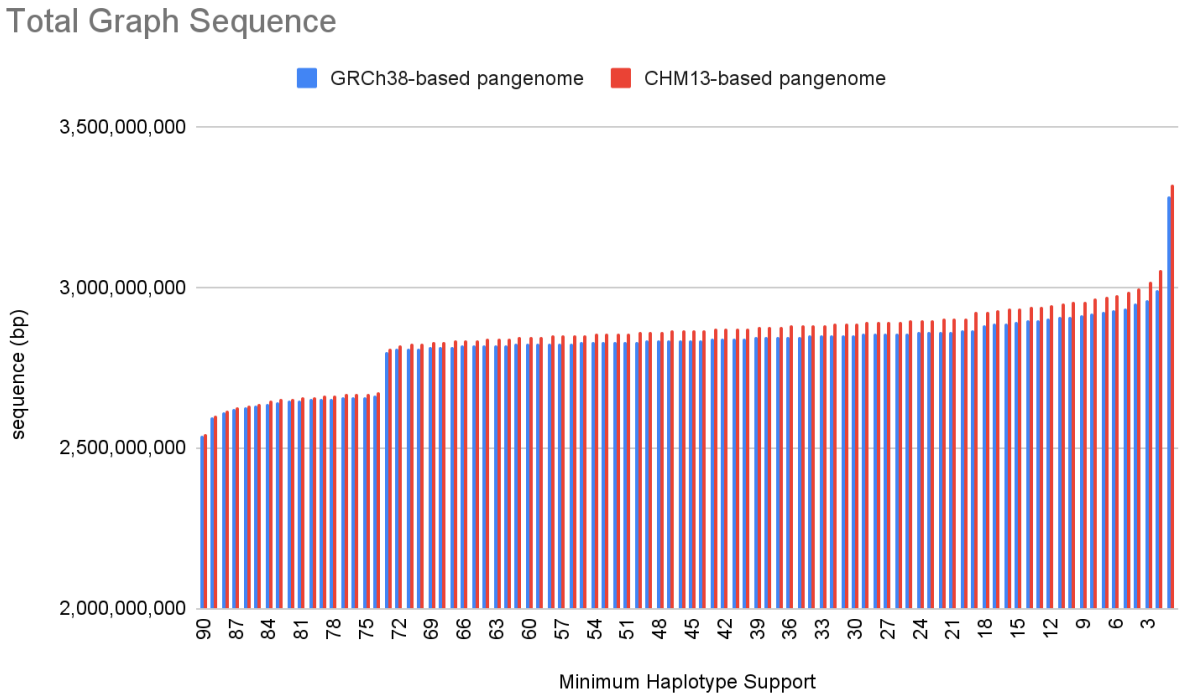
Haley J. Abel[1],Lucinda L Antonacci-Fulton[2], Mobin Asri[3], Gunjan Baid[4], Carl A. Baker[5], Anastasiya Belyaeva[4], Konstantinos Billis[6], Guillaume Bourque[7,8,9], Silvia Buonaiuto[10], Andrew Carroll[4], Mark JP Chaisson[11], Pi-Chuan Chang[4], Xian H. Chang[3], Haoyu Cheng[12,13], Justin Chu[12], Sarah Cody[2], Vincenza Colonna[10,14], Daniel E. Cook[4], Robert M. Cook-Deegan[15], Omar E. Cornejo[16], Mark Diekhans[3], Daniel Doerr[17], Peter Ebert[17], Jana Ebler[17], Evan E. Eichler[5,18], Jordan M. Eizenga[3], Susan Fairley[6], Olivier Fedrigo[19], Adam L. Felsenfeld[20], Xiaowen Feng[12,13], Christian Fischer[14], Paul Flicek[6], Giulio Formenti[19], Adam Frankish[6], Robert S. Fulton[2], Yan Gao[21], Shilpa Garg[22], Erik Garrison[14], Carlos Garcia Giron[6], Richard E. Green[23,24], Cristian Groza[25], Andrea Guarracino[26], Leanne Haggerty[6], Ira Hall[27,28], William T Harvey[5], Marina Haukness[3], David Haussler[3,18], Simon Heumos[29,30], Glenn Hickey[3], Kendra Hoekzema[5], Thibaut Hourlier[6], Kerstin Howe[31], Miten Jain[32], Erich D. Jarvis[33,18], Hanlee P. Ji[34], Alexey Kolesnikov[4], Jan O. Korbel[35], Jennifer Kordosky[5], Sergey Koren[36], HoJoon Lee[34], Alexandra P. Lewis[5], Heng Li[12,13], Wen-Wei Liao[2,37,27], Shuangjia Lu[27], Tsung-Yu Lu[38], Julian K. Lucas[3], Hugo Magalhães[17], Santiago Marco-Sola[39,40], Pierre Marijon[17], Charles Markello[3], Tobias Marschall[17], Fergal J. Martin[6], Ann McCartney[36], Jennifer McDaniel[41], Karen H. Miga[3], Matthew W. Mitchell[42], Jean Monlong[3], Jacquelyn Mountcastle[19], Katherine M. Munson[5], Moses Njagi Mwaniki[43], Maria Nattestad[4], Adam M. Novak[3], Sergey Nurk[36], Hugh E. Olsen[3], Nathan D. Olson[41], Benedict Paten[3], Trevor Pesout[3], Adam M. Phillippy[36], Alice B. Popejoy[44], David Porubsky[5], Pjotr Prins[14], Daniela Puiu[45], Mikko Rautiainen[36], Allison A Regier[2], Arang Rhie[36], Samuel Sacco[46], Ashley D. Sanders[47], Valerie A. Schneider[48], Baergen I. Schultz[20], Kishwar Shafin[4], Jonas A. Sibbesen[49], Jouni Sirén[3], Michael W. Smith[20], Heidi J. Sofia[20], Ahmad N. Abou Tayoun[50,51], Françoise Thibaud-Nissen[48], Chad Tomlinson[2], Francesca Floriana Tricomi[6], Flavia Villani[14],

Mitchell R. Vollger[5,52], Justin Wagner[41], Brian Walenz[36], Ting Wang[53], Jonathan M. D. Wood[31], Aleksey V. Zimin[45,54], Justin M. Zook[41]

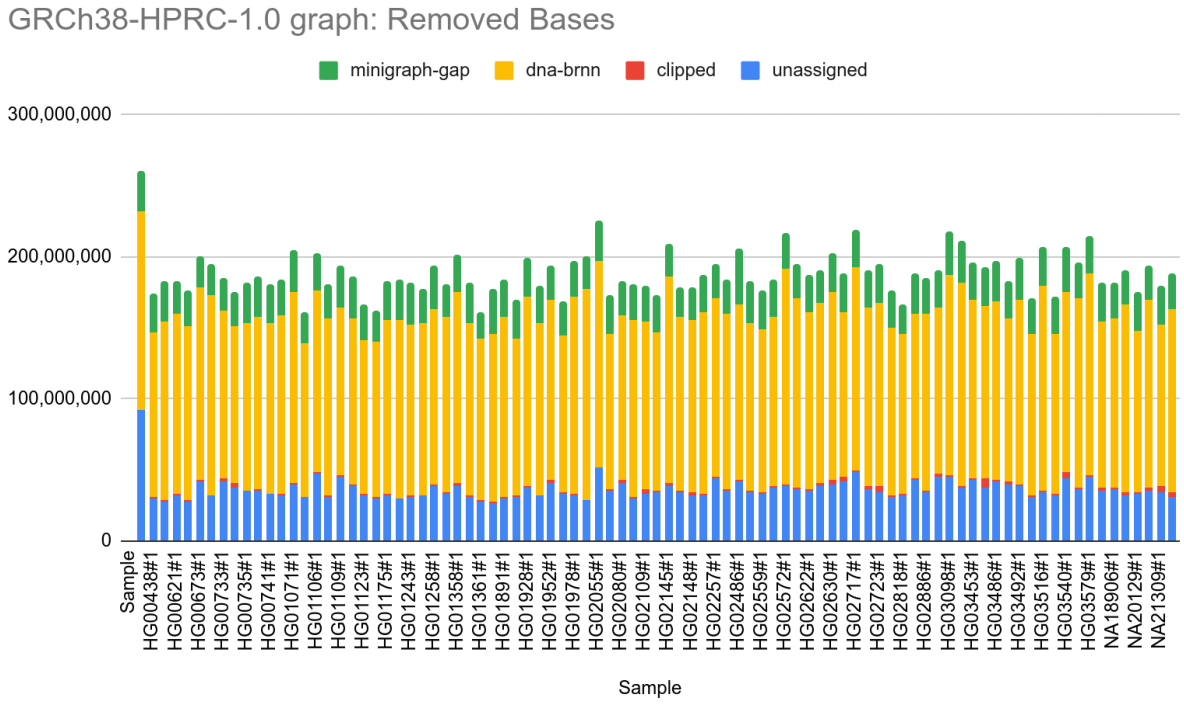
1 Division of Oncology, Department of Internal Medicine, Washington University School of Medicine, St. Louis, MO 63110, USA
2 McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO 63108, USA
3 UC Santa Cruz Genomics Institute, University of California, Santa Cruz, 1156 High St, Santa Cruz, CA, USA
4 Google LLC, 1600 Amphitheater Pkwy, Mountain View, CA 94043, USA
5 Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA 98195, USA
6 European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Cambridge, CB10 1SD, UK
7 Department of Human Genetics, McGill University, Montreal, Québec H3A 0C7, Canada
8 Canadian Center for Computational Genomics, McGill University, Montreal, Québec H3A 0G1, Canada
9 Institute for the Advanced Study of Human Biology (WPI-ASHBi), Kyoto University, Kyoto 606-8501, Japan
10 Institute of Genetics and Biophysics, National Research Council, Naples 80111, Italy
11 University of Southern California, Quantitative and Computational Biology, Los Angeles, CA, USA
12 Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA 02215, USA
13 Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02215, USA
14 Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, TN 38163, USA
15 Arizona State University, Barrett & O'Connor Washington Center, Washington DC, USA
16 School of Biological Sciences, Washington State University, Pullman WA 99163, USA
17 Institute for Medical Biometry and Bioinformatics, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany
18 Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA
19 The Vertebrate Genome Laboratory, The Rockefeller University, New York, NY 10065, USA
20 National Institutes of Health (NIH)–National Human Genome Research Institute, Bethesda, MD, USA
21 Center for Computational and Genomic Medicine, The Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA.
22 Department of Biology, University of Copenhagen, Denmark
23 Department of Biomolecular Engineering, University of California, Santa Cruz, 1156 High St., Santa Cruz, CA 95064, USA
24 Dovetail Genomics, Scotts Valley, CA 95066, USA
25 Quantitative Life Sciences, McGill University, Montreal, Québec H3A 0C7, Canada
26 Genomics Research Centre, Human Technopole, Milan 20157, Italy
27 Department of Genetics, Yale University School of Medicine, New Haven, CT 06510, USA
28 Center for Genomic Health, Yale University School of Medicine, New Haven, CT 06510, USA
29 Quantitative Biology Center (QBiC), University of Tübingen, Tübingen 72076, Germany
30 Biomedical Data Science, Department of Computer Science, University of Tübingen, Tübingen 72076, Germany
31 Tree of Life, Wellcome Sanger Institute, Hinxton, Cambridge, CB10 1SA, UK
32 Northeastern University, Boston, MA 02115, USA
33 The Rockefeller University, New York, NY 10065, USA
34 Division of Oncology, Department of Medicine, Stanford University School of Medicine, Stanford, CA, 94305, USA

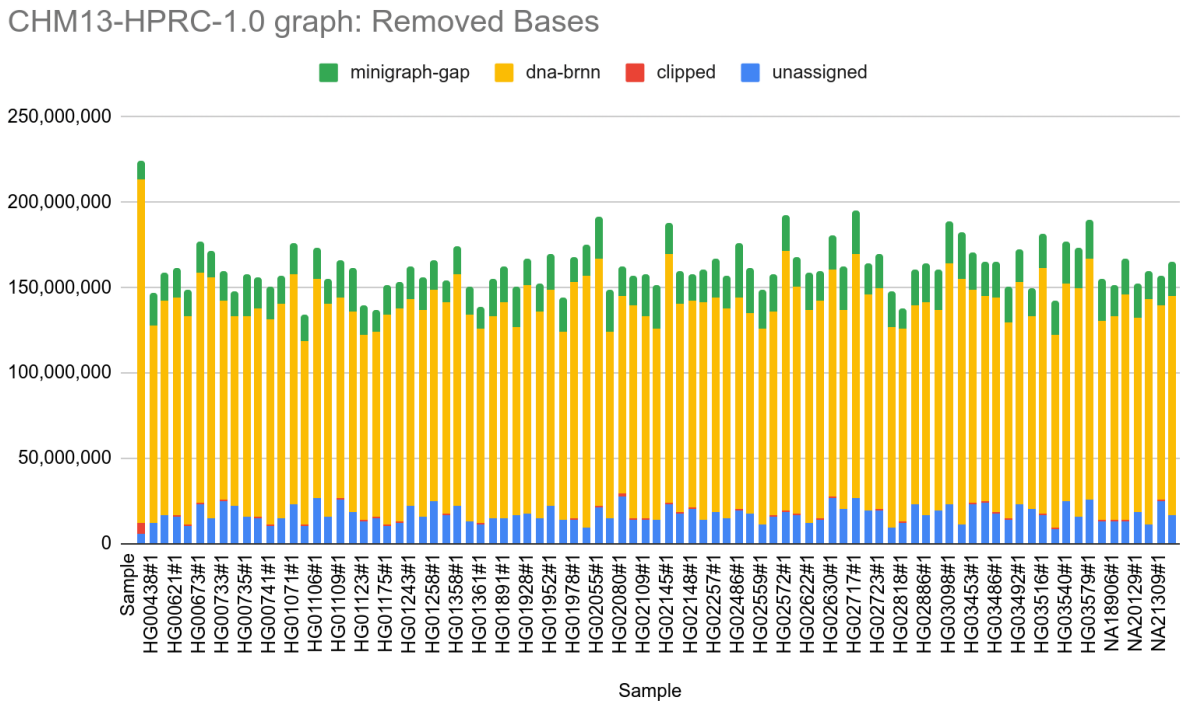
35 European Molecular Biology Laboratory, Genome Biology Unit, Meyerhofstr. 1, 69117 Heidelberg, Germany
36 Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA
37 Department of Medicine, Washington University School of Medicine, St. Louis, MO 63110, USA
38 University of Southern California, Quantitative and Computational Biology, 3551 Trousdale, Pkwy, Los Angeles, CA, USA
39 Computer Sciences Department, Barcelona Supercomputing Center, Barcelona, Spain
40 Departament d'Arquitectura de Computadors i Sistemes Operatius, Universitat Autònoma de Barcelona, Barcelona, Spain
41 Material Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, MD 20877, USA
42 Coriell Institute for Medical Research, Camden, NJ 08103, USA
43 Department of Computer Science, University of Pisa, Pisa 56127, Italy
44 Department of Public Health Sciences, University of California, Davis, One Shields Avenue, Medical Sciences 1C, Davis, CA 95616
45 Department of Biomedical Engineering, Johns Hopkins University, Baltimore 21218, MD, USA
46 Department of Ecology & Evolutionary Biology, University of California, Santa Cruz, 1156 High St, Santa Cruz, CA, USA
47 Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association, Berlin, Germany
48 National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA
49 Center for Health Data Science, University of Copenhagen, Denmark
50 Al Jalila Genomics Center of Excellence, Al Jalila Children's Specialty Hospital, Dubai, UAE
51 Center for Genomic Discovery, Mohammed Bin Rashid University of Medicine and Health Sciences, Dubai, UAE
52 Division of Medical Genetics, University of Washington School of Medicine, Seattle, WA 98195, USA
53 Department of Genetics, Washington University School of Medicine, St. Louis, MO 63110, USA
54 Center for Computational Biology, Johns Hopkins University, Baltimore, MD 21218, USA

Supplement

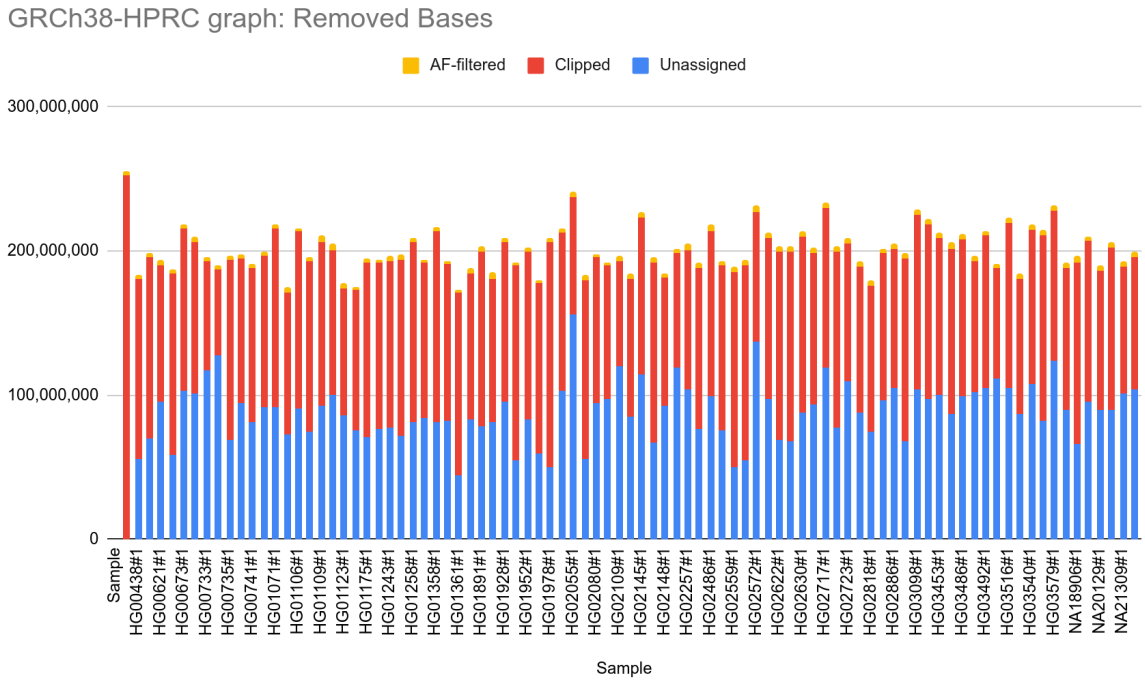


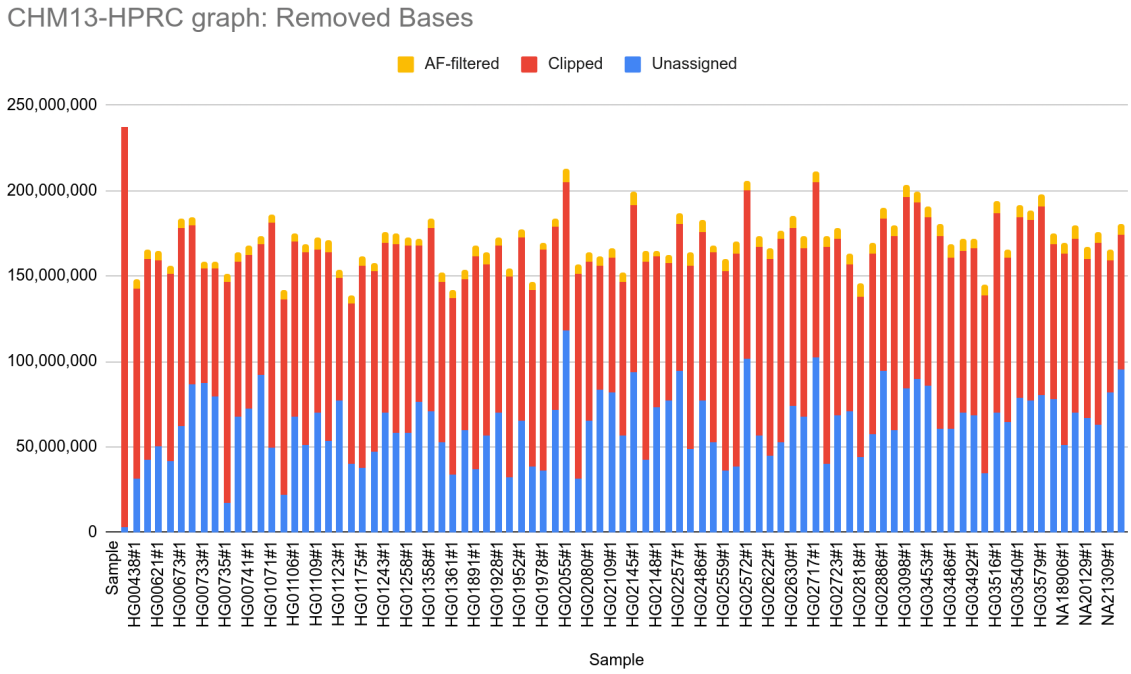
Supplementary Figure 1: The amount of sequence in the HPRC graphs by the minimum number of haplotypes that contain it. The step in the graph is due to 14 male haplotypes not possessing an X chromosome.



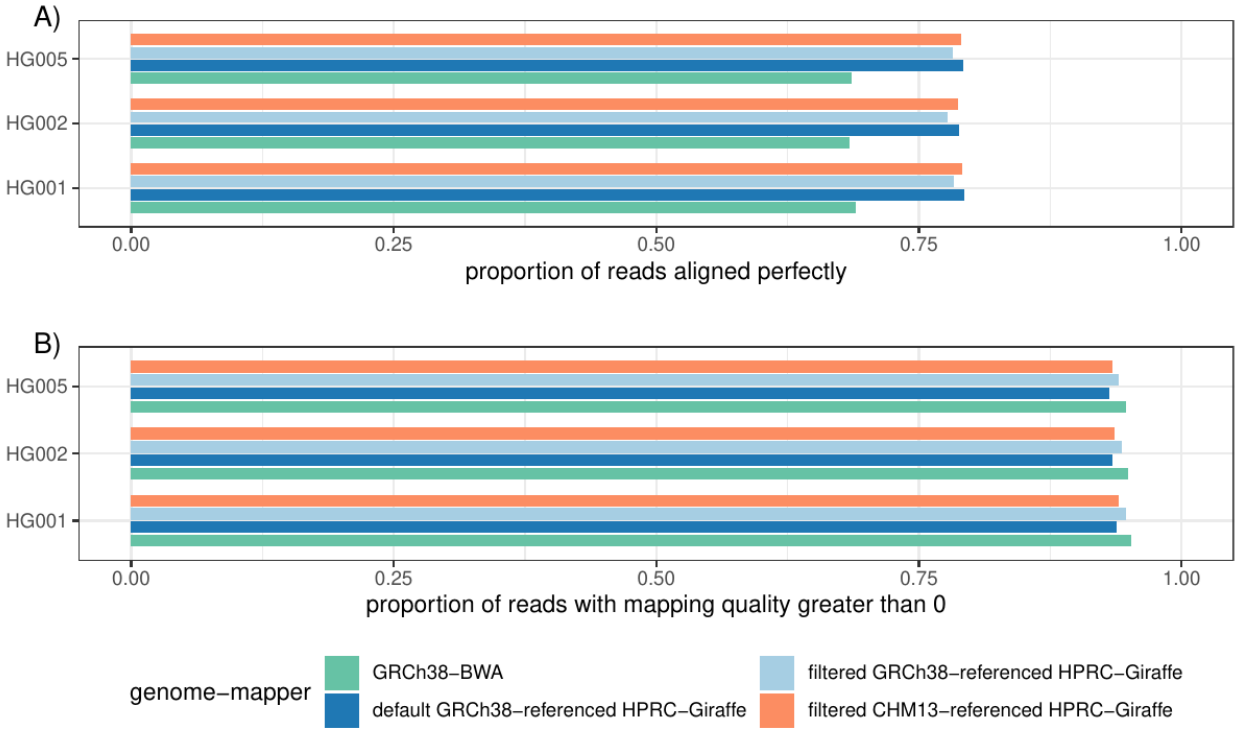


Supplementary Figure 2: Sequence excluded from the HPRC pangenomes.

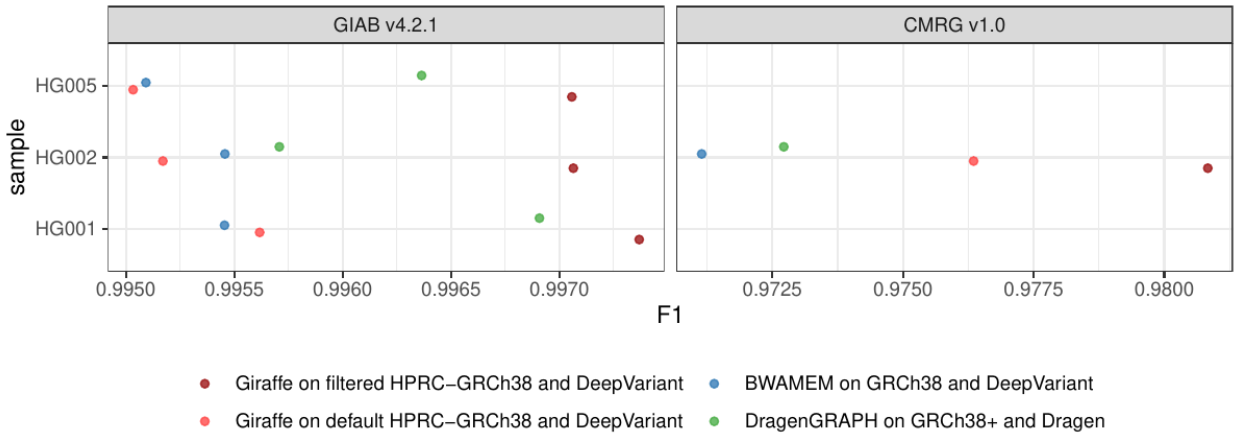




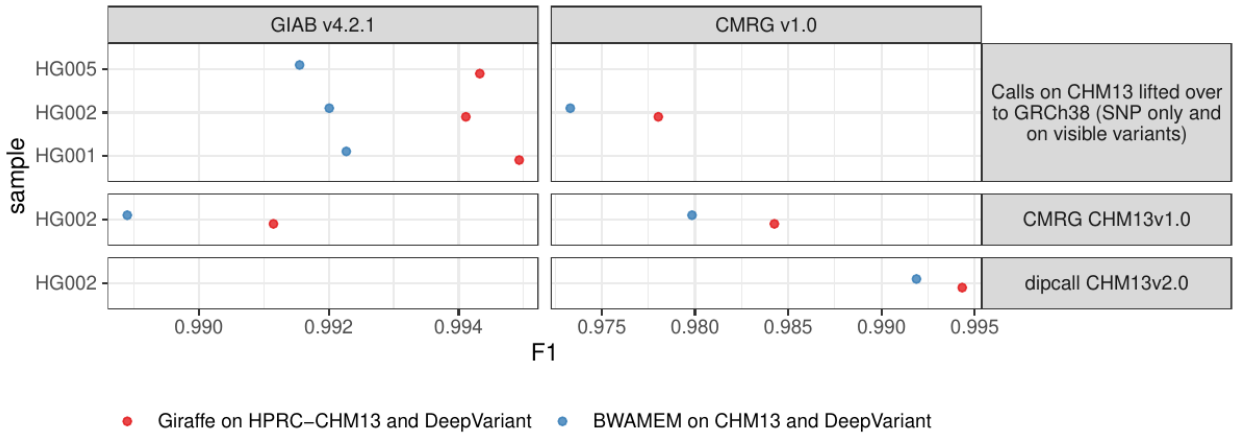
Supplementary Figure 3: Sequence excluded from the HPRC pangenomes when using the current pipeline (without dna-brnn preprocessing).



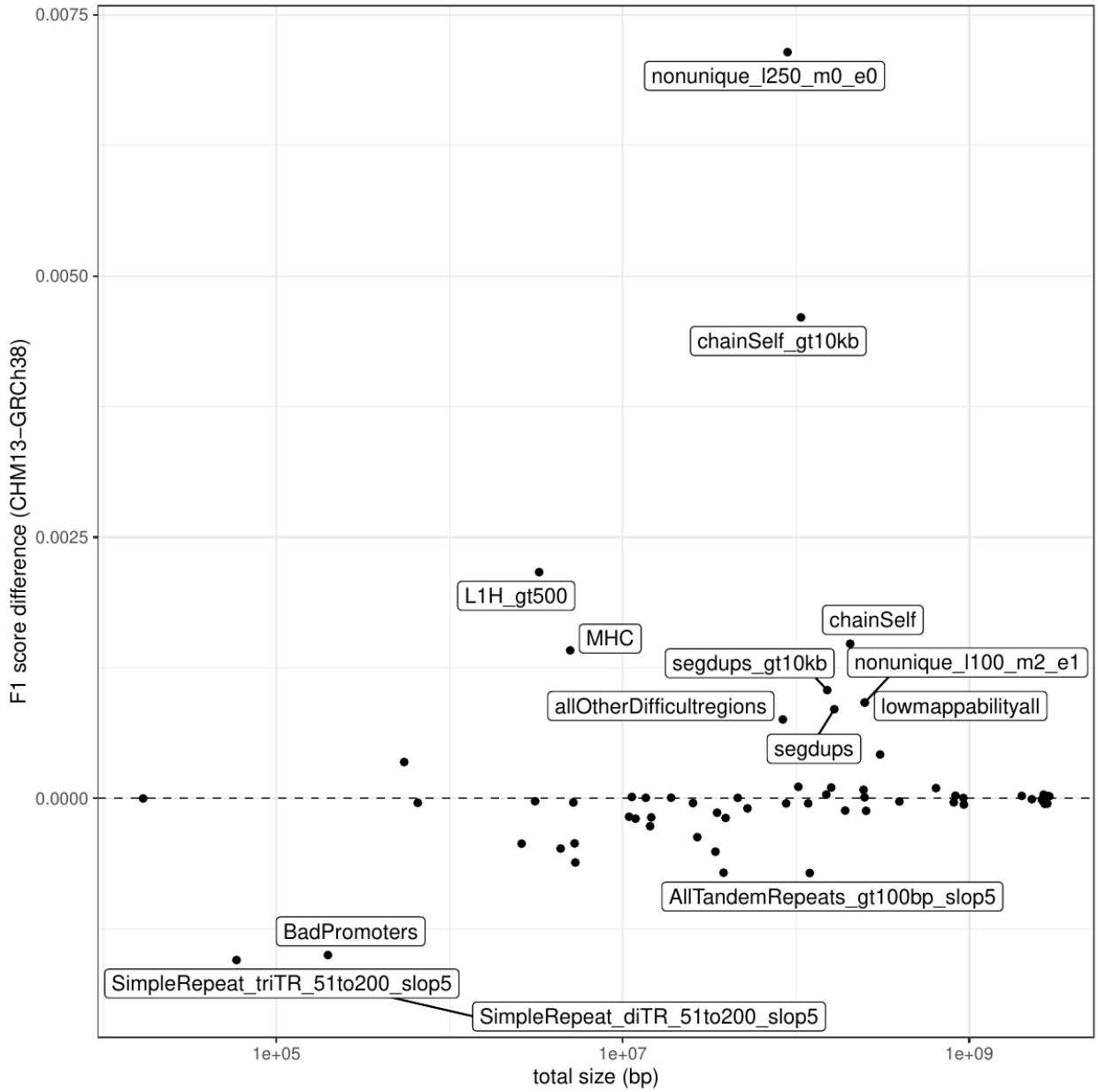
Supplementary Figure 4*:* ~30x Illumina short-reads for three GIAB samples were mapped using three approaches: BWAMEM on GRCh38 (red), vg Giraffe on the GRCh38-referenced HPRC pangenome (blue) and vg Giraffe on the CHM13-referenced pangenome (green). The darker blue bar corresponds to the default GRCh38-based HPRC pangenome, while the lighter blue to the frequency-filtered pangenome used in practice for read mapping and variant calling. A) Proportion of the reads aligning perfectly to the (pan-)genome for each sample (y-axis). B) Proportion of reads with a mapping quality greater than 0.



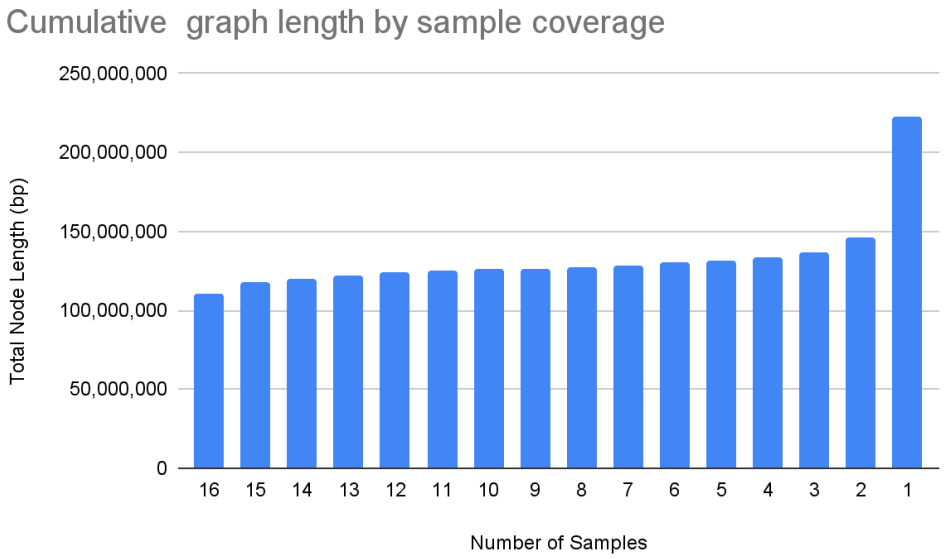
Supplementary Figure 5*:* Evaluation of calls made on both the default pangenome (light red) and the frequency-filtered pangenome (dark red). The results when aligning reads with BWAMEM (blue) or using the Dragen pipeline (green) are also shown. The F1 score is shown on the x-axis across samples from the Genome in a Bottle (y-axis). Left: Genome in a Bottle v4.2.2 truth set. Right: Challenging Medically Relevant Genes v1.0 truth set.



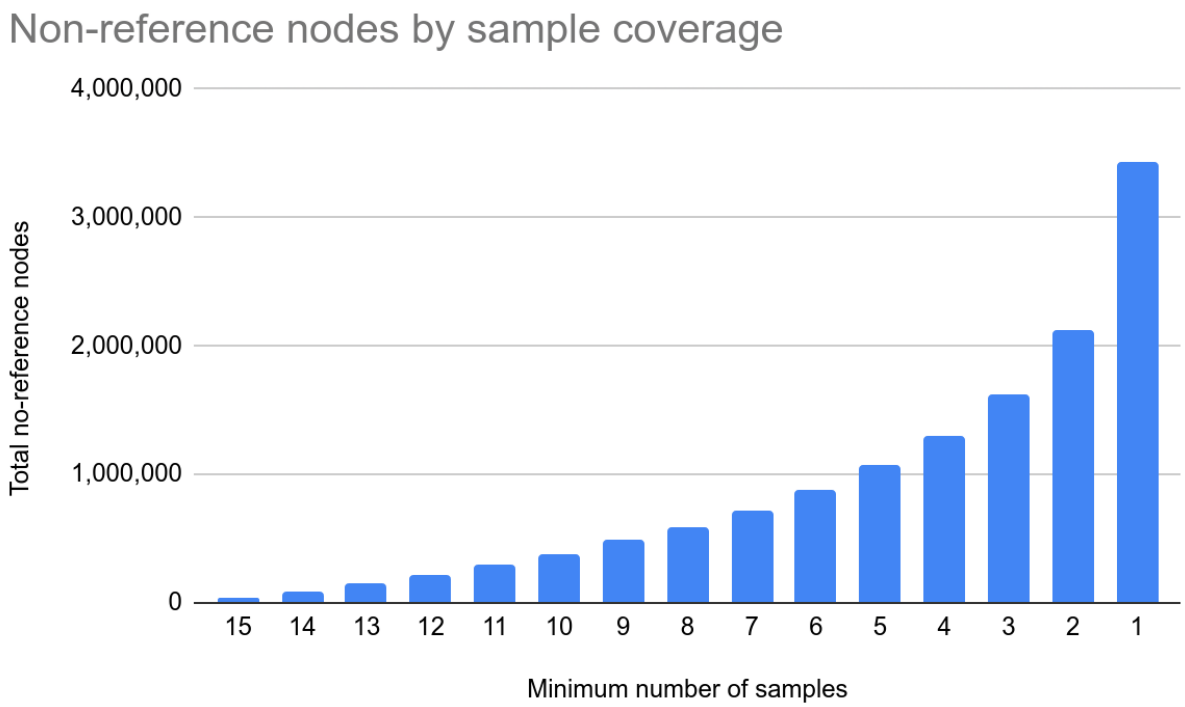
Supplementary Figure 6: Evaluation of calls made on CHM13: aligning reads with BWAMEM (blue), or to the CHM13-based HPRC pangenome and projecting them to CHM13 (red). The F1 score is shown on the x-axis across samples from the Genome in a Bottle (y-axis). Left: Genome in a Bottle v4.2.2 truth set. Right: Challenging Medically Relevant Genes v1.0 truth set. Three approaches are shown as horizontal panels. Top: variants called on CHM13 were lifted over to be evaluated against the GRCh38 truth sets. Only SNPs and variant that are visible (not homozygous for the reference allele) on both reference genomes were used. Middle: the CMRG truth set for CHM13 v1.0 was lifted to CHM13 v2.0. The whole genome evaluation (left) was limited to the GIAB v4.2.1 confident regions lifted from GRCh38 to CHM13. Bottom: Preliminary draft truth set for CHM13 v2.0 based on HiFi assemblies analyzed with dipcall.



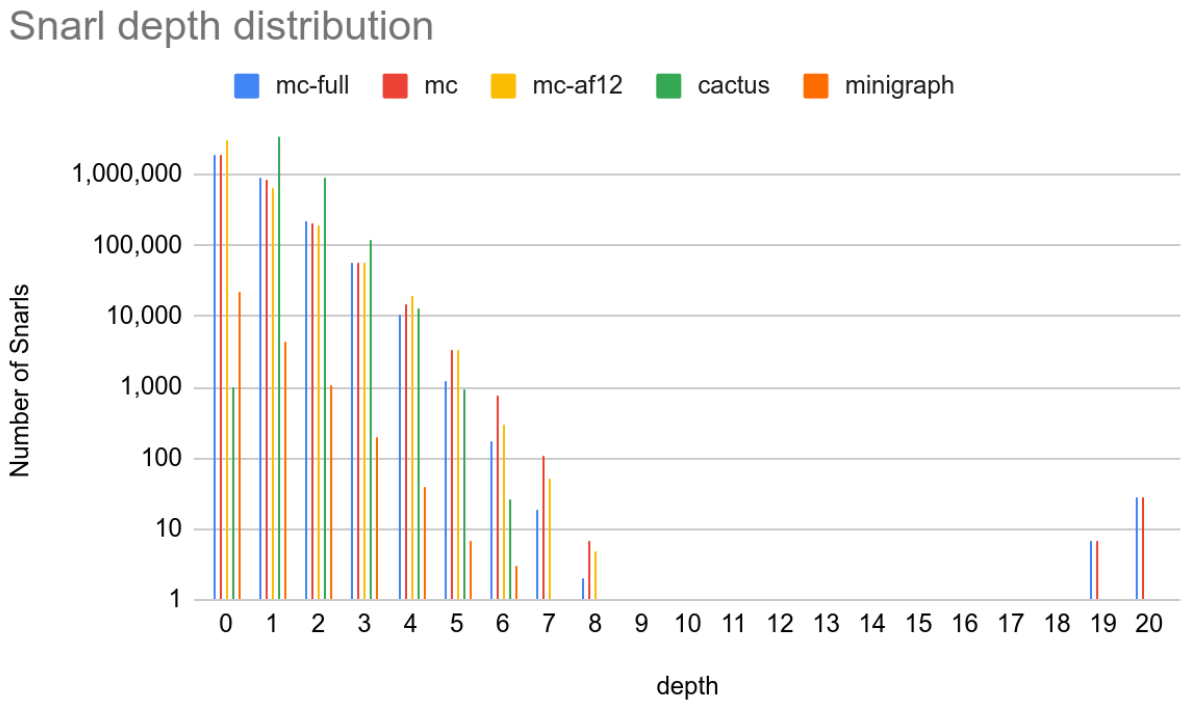
Supplementary Figure 7: Difference between the F1 score obtained when using the CHM13-based pangenome compared to the GRCh38-based pangenome (y-axis), stratified by region sets from the GIAB (points). The total amount of sequence that represents each region set is shown on the x-axis. The top 10 most regions with the largest differences are labeled.



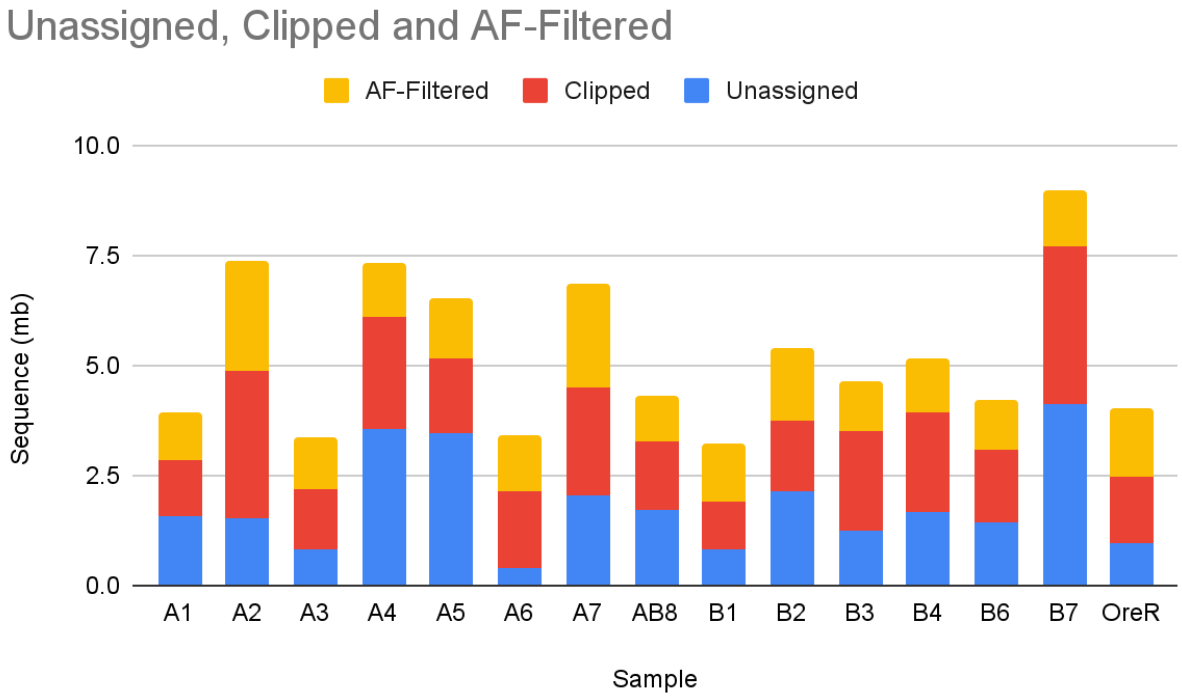
Supplementary Figure 8: The amount of sequence in the *D. melanogaster* graph by the minimum number of haplotypes that contain it.



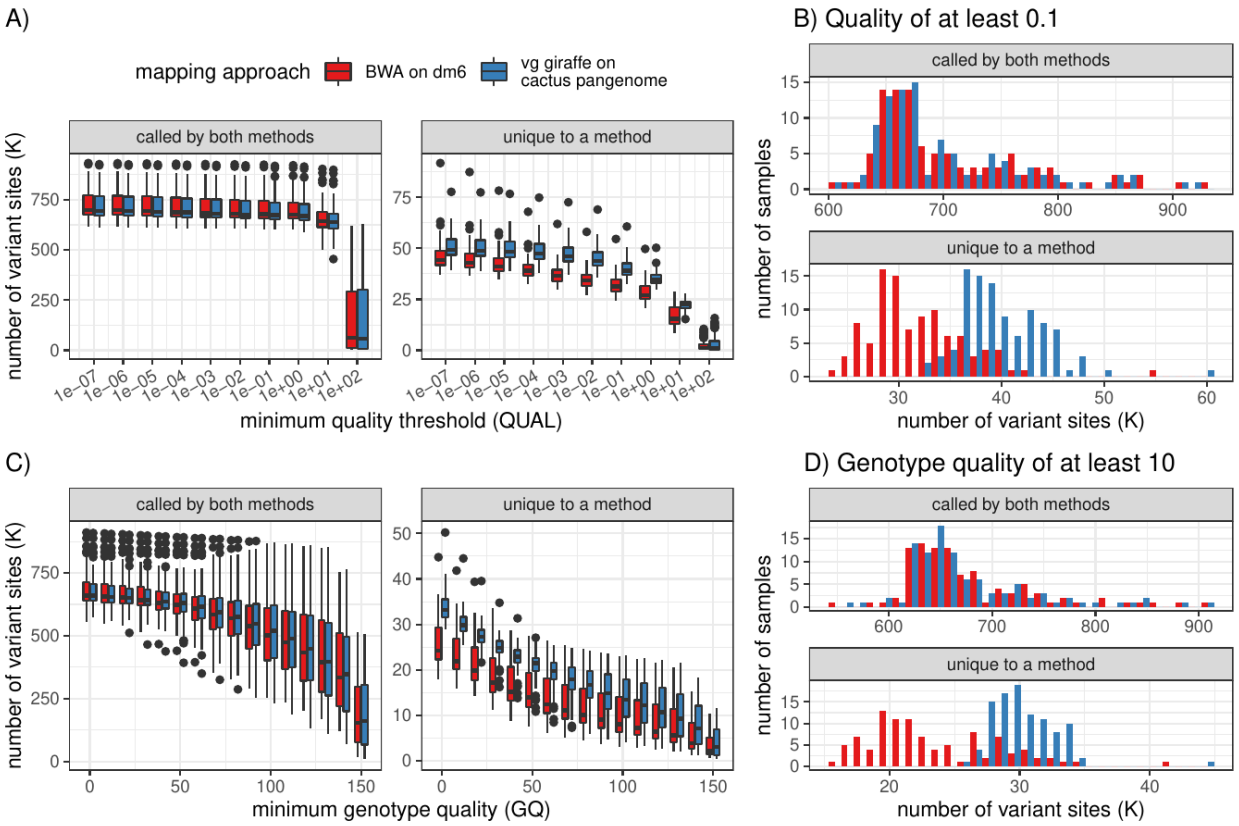
Supplementary Figure 9: The number of nodes not present in dm6 covered by at least the given number of samples.



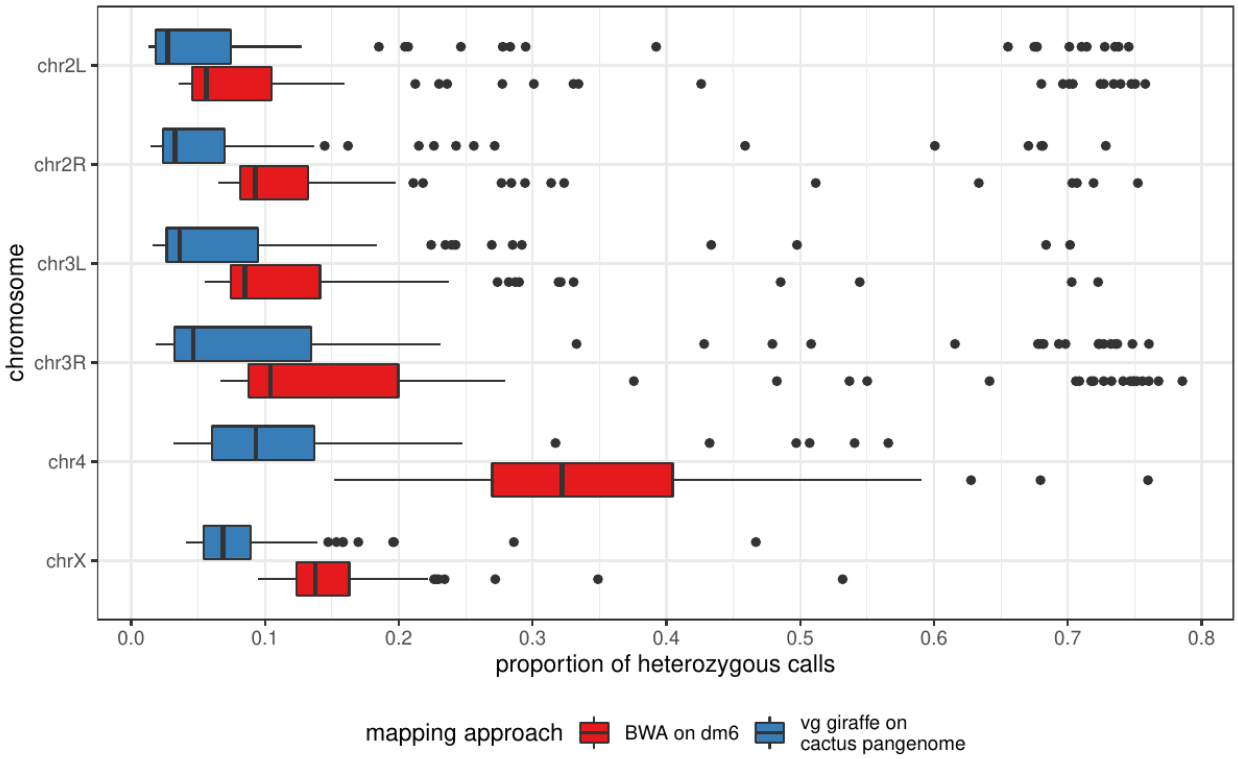
Supplementary Figure 10: Snarl depth distribution.



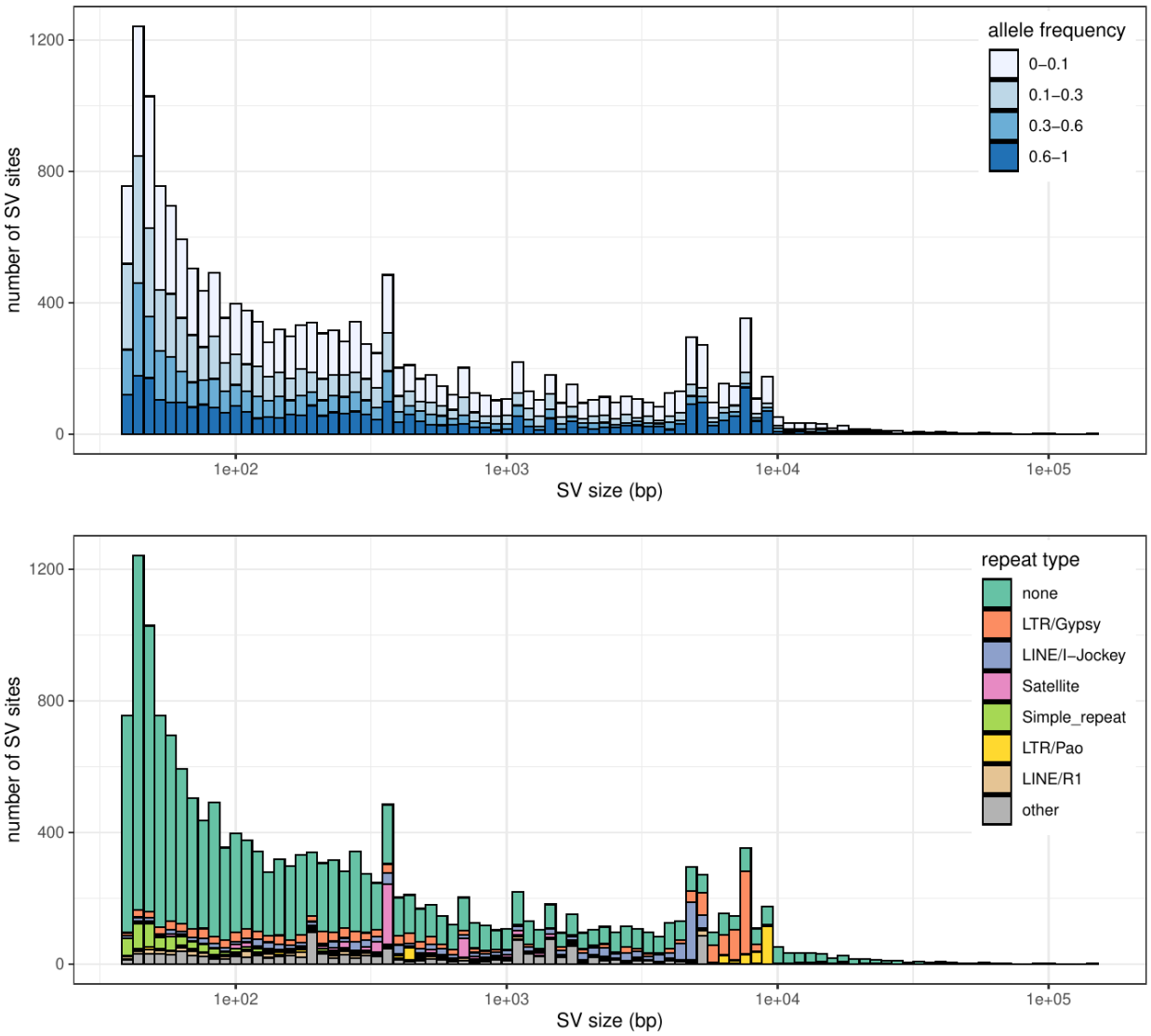
Supplementary Figure 11: Sequence excluded from the *D. melanogaster* pangenome.



Supplementary Figure 12: Number of variant sites with an alternate allele called in each of the 100 samples with FreeBayes. Two mapping approaches are compared: short-reads mapped to dm6 using BWA-MEM (red); short-reads mapped to the pangenome using vg Giraffe (blue). The variant sites were split into sites found by both approaches and sites found only by one. **A)** Distribution of the number of variant sites for different minimum quality (QUAL field) (x-axis). **B)** Only variant sites with a quality of at least 0.1 were counted. This corresponds to x=0.1 in A). **C)** Distribution of the number of variant sites for different minimum genotype quality (GQ field) (x-axis). **D)** Only variant sites with a genotype quality of at least 10 were counted. This corresponds to x=10 in C).



Supplementary Figure 13: Proportion of heterozygous small variants called by FreeBayes in each of the 100 fly samples (point). Reads were either aligned to the pangenome and projected to dm6 (blue), or mapped to dm6 with BWA-MEM (red). Due to the inbreeding of these lines, we expect low heterozygosity.



Supplementary Figure 14: Distribution of the size of the SVs genotyped across 100 fly samples. The x-axis is log-scaled. **Top:** The SVs are colored by their allele frequencies. **Bottom:** The SVs are colored by the repeat class as annotated by Repeat Masker (Smit et al., 2013-2015)

Graph	Nodes	Edges	Total Node Length	Total Non-ref Node Length	Total Path Length
SV Graph (GRCh38)	424,643	637,628	3,239,764,787	140,014,069	3,239,764,787
SV Graph (CHM13)	493,631	738,529	3,365,688,482	253,629,026	3,365,688,482
GRCh38-based Pangenome	81,751,614	113,258,931	3,287,932,785	188,182,067	254,821,009,311
GRCh38-based Filtered Pangenome (AF>=10%)	59,960,908	72,408,601	3,153,443,019	53,692,301	254,415,272,646
CHM13-based Pangenome	85,591,995	118,409,526	3,324,657,754	212,598,298	257,143,252,360
CHM13-based Filtered Pangenome (AF>=10%)	62,335,399	75,270,997	3,166,744,316	54,684,860	256,673,009,341

Supplementary Table 1: HPRC graph sizes

Graph	Nodes	Edges	Total Node Length	Total Path Length
SV Graph	80,853	112,742	214,547,800	214,547,800
Unclipped Pangenome	9,042,502	12,364,039	251,857,504	2,182,961,082
Pangenome	8,978,195	12,276,452	223,071,144	2,152,888,069
Filtered Pangenome (AF12.5%)	7,686,219	9,788,690	202,497,872	2,131,677,729
Progressive Cactus Graph	12,974,720	17,684,675	470,148,493	2,216,588,031

Supplementary Table 2: *D. Melegonaster* graph sizes

Phase	dm6	hprc-v1.0 grch38	hprc-v1.0 chm13	hprc grch38 (new pipeline)	hprc chm13 (new pipeline)
Minigraph construction	3.02	45.43	39.82	45.82	21.4
Minigraph mapping	1.01	3.14	4.52	1.93	1.58
Split by chromosome	0.11	1.6	2.1	1.67	1.43
Cactus alignment	0.43	11.56	7.66	5.04	5.27
Indexing and clipping (full graph)	0.13	N/A	N/A	4.37	4.68
Indexing and clipping (clipped graph)	0.26	10.02	10.98	3.17	4.32
Indexing and clipping (AF12.5% graph)	0.16	10.81*	10.75*	3.42	4.23
Total	5.12	71.75	75.83	65.42	42.91

Supplementary Table 3: Minigraph-Cactus running times. The “new pipeline” columns refer to graphs made using the method described here which does not rely on dna-brnn for clipping. The dm6 graphs were made using up to 32 cores and 16Gb RAM. The HPRC graphs were made on an AWS cluster using up to 25 32 core 256Gb RAM machines, except for the indexing stages which were done on up to 2 64 core 512Gb RAM machines.
** These values were not kept in the logs and were estimated using the ratios in the neighboring columns (ex 10.81 = 3.42/3.17 * 10.02).*

Phase	dm6
Lastz repeatmasking	0.38
All-to-all lastz	17.97

alignment	
Cactus alignment	0.83
Total	19.18

Supplementary Table 4: Progressive Cactus running times using single 32-core machine with up to 64Gb RAM.

DGRP Line	Sequencing Technology	Freeze	Mapped Coverage	Raw Read Length:Read Number	NCBI SRA	NCBI SRR
DGRP_21	Illumina	F1	15.8	95bp:37046984	SRX021040	SRR834526
DGRP_31	Illumina	F2	49.2	125bp:76894692	SRX155996	SRR834509
DGRP_32	Illumina	F2	56.2	125bp:88154526	SRX155997	SRR834512
DGRP_38	Illumina	F1	28.0	95bp:56154204	SRX025317	SRR834541
DGRP_40	Illumina	F1	33.3	95bp:69063428	SRX021235	SRR835025
DGRP_42	Illumina	F1	20.2	95bp:37186556	SRX021255	SRR835027
DGRP_48	Illumina	F2	32.7	125bp:58419132	SRX155989	SRR835034
DGRP_49	Illumina	F1	15.2	75bp:37870818	SRX021267	SRR835037
DGRP_57	Illumina	F1	32.6	100bp:64966990	SRX021296	SRR933581
DGRP_75	Illumina	F1	18.5	110bp:38161744	SRX021384	SRR835087
DGRP_83	Illumina	F1	16.3	75bp:41070470	SRX023456	SRR835058
DGRP_100	Illumina	F2	52.3	125bp:87340978	SRX156026	SRR833244
DGRP_138	Illumina	F1	30.1	100bp:61689820	SRX021008	SRR932121
DGRP_142	Illumina	F1	19.7	110bp:41167794	SRX020759	SRR834551
DGRP_177	Illumina	F1	24.6	95bp:49114764	SRX021026	SRR834547
DGRP_181	Illumina	F1	24.7	75bp:64093862	SRX020912	SRR933563
DGRP_189	Illumina	F2	37.8	125bp:63289120	SRX155979	SRR834523
DGRP_223	Illumina	F2	40.8	125bp:71152512	SRX155994	SRR834527
DGRP_235	Illumina	F1	18.4	95bp:38296004	SRX021053	SRR834531
DGRP_318	Illumina	F1	15.2	75bp:39068236	SRX021082	SRR834507

DGRP_319	Illumina	F2	37.6	125bp:70621686	SRX155981	SRR834508
DGRP_320	Illumina	F1	24.2	95bp:51875680	SRX021063	SRR834510
DGRP_321	Illumina	F1	33.5	95bp:67314152	SRX021094	SRR834511
DGRP_332	Illumina	F1	25.7	75bp:65583082	SRX021095	SRR933569
DGRP_348	Illumina	F2	48.3	125bp:78515972	SRX156029	SRR834514
DGRP_352	Illumina	F1	15.6	75bp:44982388	SRX021101	SRR834516
DGRP_354	Illumina	F2	57.2	101bp:106369344	SRX156027	SRR834517
DGRP_355	Illumina	F2	44.9	101bp:84541222	SRX156028	SRR834545
DGRP_356	Illumina	F1	15.5	75bp:42903612	SRX023833	SRR834537
DGRP_359	Illumina	F1	20.2	95bp:37271884	SRX023424	SRR834546
DGRP_361	Illumina	F2	40.6	125bp:68254340	SRX155984	SRR834553
DGRP_370	Illumina	F1	20.9	95bp:43793604	SRX021104	SRR834539
DGRP_377	Illumina	F1	21.8	95bp:43796182	SRX023834	SRR834543
DGRP_381	Illumina	F1	20.9	75bp:54335852	SRX021112	SRR933573
DGRP_382	Illumina	F2	41.1	125bp:73812254	SRX156013	SRR834552
DGRP_383	Illumina	F1	19.1	95bp:39897030	SRX021113	SRR834554
DGRP_390	Illumina	F2	26.2	125bp:42709922	SRX156014	SRR834519
DGRP_392	Illumina	F1	23.2	95bp:51156860	SRX021157	SRR834520
DGRP_395	Illumina	F2	47.1	101bp:87233368	SRX156015	SRR834521
DGRP_397	Illumina	F2	30.0	125bp:48910026	SRX156017	SRR834522
DGRP_405	Illumina	F1	22.9	95bp:50080536	SRX021242	SRR835023
DGRP_406	Illumina	F1	25.0	95bp:51821248	SRX021254	SRR835024
DGRP_426	Illumina	F1	21.1	95bp:43746634	SRX021245	SRR835026
DGRP_427	Illumina	F1	16.3	45bp:64106936	SRX006155	SRR933577
DGRP_439	Illumina	F1	20.4	95bp:44762436	SRX021244	SRR835028
DGRP_440	Illumina	F1	17.2	95bp:43161850	SRX021246	SRR835029
DGRP_441	Illumina	F1	18.7	95bp:42278010	SRX023835	SRR835030
DGRP_443	Illumina	F1	28.5	95bp:57567568	SRX021260	SRR835031
DGRP_461	Illumina	F1	21.9	95bp:49324528	SRX021262	SRR835033

DGRP_491	Illumina	F1	15.1	75bp:40944392	SRX021268	SRR835035
DGRP_492	Illumina	F1	22.1	95bp:44580310	SRX021270	SRR835036
DGRP_502	Illumina	F1	21.7	95bp:44336646	SRX021271	SRR835038
DGRP_505	Illumina	F2	43.7	125bp:71295212	SRX156002	SRR835039
DGRP_508	Illumina	F1	21.2	95bp:42338556	SRX021272	SRR835040
DGRP_509	Illumina	F1	15.3	75bp:38095912	SRX021273	SRR835041
DGRP_513	Illumina	F1	19.6	95bp:42640722	SRX021282	SRR835042
DGRP_528	Illumina	F2	36.2	125bp:57697778	SRX155985	SRR835043
DGRP_530	Illumina	F2	20.7	125bp:34726088	SRX156031	SRR835044
DGRP_531	Illumina	F1	17.9	95bp:41560152	SRX021290	SRR835045
DGRP_535	Illumina	F1	15.2	75bp:40234802	SRX021293	SRR835046
DGRP_551	Illumina	F2	21.4	125bp:35225968	SRX156034	SRR835047
DGRP_555	Illumina	F1	19.2	75bp:50103810	SRX006159	SRR933580
DGRP_559	Illumina	F2	24.2	125bp:36482062	SRX156032	SRR835048
DGRP_566	Illumina	F2	48.8	101bp:89414580	SRX156033	SRR835050
DGRP_596	Illumina	F2	41.1	101bp:73915046	SRX156004	SRR835096
DGRP_627	Illumina	F2	36.7	125bp:82297368	SRX155988	SRR835097
DGRP_630	Illumina	F2	21.7	125bp:36162916	SRX156003	SRR835098
DGRP_634	Illumina	F2	19.4	125bp:32632568	SRX156018	SRR835086
DGRP_705	Illumina	F1	16.7	75bp:47006608	SRX006162	SRR933585
DGRP_707	Illumina	F1	17.8	75bp:46657404	SRX006163	SRR933586
DGRP_712	Illumina	F1	16.3	75bp:44687868	SRX006164	SRR933587
DGRP_727	Illumina	F1	27.5	75bp:73781476	SRX021382	SRR933589
DGRP_732	Illumina	F1	16.3	75bp:42170344	SRX006167	SRR933591
DGRP_737	Illumina	F1	25.1	75bp:74740132	SRX023451	SRR933592
DGRP_738	Illumina	F1	27.1	75bp:75804508	SRX021383	SRR933593
DGRP_757	Illumina	F1	28.4	75bp:74326240	SRX021385	SRR933594
DGRP_761	Illumina	F1	15.2	75bp:40867250	SRX021386	SRR835088
DGRP_776	Illumina	F1	15.6	75bp:39890986	SRX021387	SRR835089

DGRP_787	Illumina	F1	15.4	75bp:39795416	SRX021388	SRR835091
DGRP_790	Illumina	F1	17.0	95bp:35620658	SRX021389	SRR835092
DGRP_805	Illumina	F1	16.1	75bp:43182102	SRX021400	SRR835095
DGRP_810	Illumina	F1	15.5	75bp:36972402	SRX021418	SRR835051
DGRP_812	Illumina	F1	16.1	75bp:38719004	SRX021419	SRR835052
DGRP_819	Illumina	F2	73.0	100bp:150745358	SRX156006	SRR835054
DGRP_822	Illumina	F1	17.7	110bp:41079524	SRX021476	SRR835055
DGRP_837	Illumina	F1	20.7	95bp:46411538	SRX021479	SRR933599
DGRP_843	Illumina	F2	42.3	125bp:68658714	SRX156036	SRR835059
DGRP_849	Illumina	F2	39.9	125bp:61687178	SRX156035	SRR835060
DGRP_850	Illumina	F2	43.6	125bp:69699750	SRX155993	SRR835061
DGRP_855	Illumina	F1	19.2	110bp:42348166	SRX021563	SRR835062
DGRP_857	Illumina	F1	20.8	110bp:42340250	SRX021492	SRR835063
DGRP_882	Illumina	F1	17.4	75bp:44722234	SRX021496	SRR835067
DGRP_887	Illumina	F1	19.5	95bp:43595728	SRX021527	SRR835069
DGRP_890	Illumina	F1	15.9	75bp:41954706	SRX021499	SRR835071
DGRP_892	Illumina	F1	20.5	95bp:45702226	SRX023838	SRR835072
DGRP_894	Illumina	F1	16.8	95bp:35128536	SRX021528	SRR835073
DGRP_897	Illumina	F1	27.0	75bp:70892788	SRX023457	SRR933601
DGRP_907	Illumina	F1	17.5	95bp:36385056	SRX021500	SRR835074
DGRP_908	Illumina	F1	19.9	95bp:39111536	SRX021501	SRR835075
DGRP_913	Illumina	F2	43.7	125bp:69250292	SRX156024	SRR835077

Supplementary Table 5: DGRP sequencing data used for *D. Melanogaster* mapping and variant calling experiments

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., Buyske, S., Matise, T. C., Muzny, D. M., Zody, M. C., Lander, E. S., Dutcher, S. K., Stitziel, N. O., & Hall, I. M. (2020). Mapping and characterization of structural variation in 17,795 human genomes. *Nature*, *583*(7814), 83–89.

Armstrong, J., Hickey, G., Diekhans, M., Fiddes, I. T., Novak, A. M., Deran, A., Fang, Q., Xie, D., Feng, S., Stiller, J., Genereux, D., Johnson, J., Marinescu, V. D., Alföldi, J., Harris, R. S., Lindblad-Toh, K., Haussler, D., Karlsson, E., Jarvis, E. D., … Paten, B. (2020). Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature*, *587*(7833), 246–251.

Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smit, A. F. A., Roskin, K. M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E. D., Haussler, D., & Miller, W. (2004). Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Research*, *14*(4), 708–715.

broadinstitute/picard. (n.d.). GitHub. Retrieved October 5, 2022, from https://github.com/broadinstitute/picard

Chakraborty, M., Emerson, J. J., Macdonald, S. J., & Long, A. D. (2019). Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. *Nature Communications*, *10*(1), 1–11.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008.

Doerr, D. (2022). GFAffix. In *GitHub*. https://github.com/marschall-lab/GFAffix/releases/tag/0.1.3

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., Dilthey, A. T., & Marschall, T. (2022). Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of

variant classes. *Nature Genetics*, *54*(4), 518–525.

Eizenga, J. M., Novak, A. M., Kobayashi, E., Villani, F., Cisar, C., Heumos, S., Hickey, G., Colonna, V., Paten, B., & Garrison, E. (2020). Efficient dynamic variation graphs. *Bioinformatics* , *36*(21), 5139–5144.

Eizenga, J. M., Novak, A. M., Sibbesen, J. A., Heumos, S., Ghaffaari, A., Hickey, G., Chang, X., Seaman, J. D., Rounthwaite, R., Ebler, J., Rautiainen, M., Garg, S., Paten, B., Marschall, T., Sirén, J., & Garrison, E. (2020). Pangenome Graphs. *Annual Review of Genomics and Human Genetics*, *21*, 139–162.

Fiddes, I. T., Armstrong, J., Diekhans, M., Nachtweide, S., Kronenberg, Z. N., Underwood, J. G., Gordon, D., Earl, D., Keane, T., Eichler, E. E., Haussler, D., Stanke, M., & Paten, B. (2018). Comparative Annotation Toolkit (CAT)-simultaneous clade and personal genome annotation. *Genome Research*, *28*(7), 1029–1038.

Gao, Y., Liu, Y., Ma, Y., Liu, B., Wang, Y., & Xing, Y. (2021). abPOA: an SIMD-based C library for fast partial order alignment using adaptive band. *Bioinformatics* , *37*(15), 2209–2211.

Garrison, E., & Marth, G. (2012). *Haplotype-based variant detection from short-read sequencing*. https://doi.org/10.48550/arXiv.1207.3907

Garrison, E., Sirén, J., Novak, A. M., Hickey, G., Eizenga, J. M., Dawson, E. T., Jones, W., Garg, S., Markello, C., Lin, M. F., Paten, B., & Durbin, R. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, *36*(9), 875–879.

Goenka, S. D., Turakhia, Y., Paten, B., & Horowitz, M. (2020, November). SegAlign: A Scalable GPU-Based Whole Genome Aligner. *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*. SC20: International Conference for High Performance Computing, Networking, Storage and Analysis, Atlanta, GA, USA. https://doi.org/10.1109/sc41405.2020.00043

Guarracino, A., Buonaiuto, S., Potapova, T., Rhie, A., Koren, S., Rubinstein, B., Fischer, C.,

Human Pangenome Reference Consortium, Gerton, J. L., Phillippy, A. M., Colonna, V., & Garrison, E. (2022). Recombination between heterologous human acrocentric chromosomes. In *bioRxiv* (p. 2022.08.15.504037). https://doi.org/10.1101/2022.08.15.504037

Harris, R. S. (2007). *Improved pairwise alignment of genomic DNA* [Doctor of Philosophy, The Pennsylvania State University]. https://search.proquest.com/openview/bc77cca0fb9390b44b9ef572fb574322/1?pq-origsite=gscholar&cbl=18750

Hickey, G., Heller, D., Monlong, J., Sibbesen, J. A., Sirén, J., Eizenga, J., Dawson, E. T., Garrison, E., Novak, A. M., & Paten, B. (2020). Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biology*, *21*(1), 35.

Hickey, G., Paten, B., Earl, D., Zerbino, D., & Haussler, D. (2013). HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics* , *29*(10), 1341–1342.

Huang, W., Massouras, A., Inoue, Y., Peiffer, J., Ràmia, M., Tarone, A. M., Turlapati, L., Zichner, T., Zhu, D., Lyman, R. F., Magwire, M. M., Blankenburg, K., Carbone, M. A., Chang, K., Ellis, L. L., Fernandez, S., Han, Y., Highnam, G., Hjelmen, C. E., … Mackay, T. F. C. (2014). Natural variation in genome architecture among 205 Drosophila melanogaster Genetic Reference Panel lines. *Genome Research*, *24*(7), 1193–1208.

Just, W. (2004). Computational Complexity of Multiple Sequence Alignment with SP-Score. *Journal of Computational Biology*, *8*(6). https://www.liebertpub.com/doi/10.1089/106652701753307511

Kille, B., Balaji, A., Sedlazeck, F. J., Nute, M., & Treangen, T. J. (2022). Multiple genome alignment in the telomere-to-telomere assembly era. *Genome Biology*, *23*(1), 1–22.

Krusche, P., Trigg, L., Boutros, P. C., Mason, C. E., De La Vega, F. M., Moore, B. L., Gonzalez-Porta, M., Eberle, M. A., Tezak, Z., Lababidi, S., Truty, R., Asimenos, G., Funke, B., Fleharty, M., Chapman, B. A., Salit, M., & Zook, J. M. (2019). Best practices for

benchmarking germline small-variant calls in human genomes. *Nature Biotechnology*, *37*(5), 555–560.

Kuhn, R. M., Haussler, D., & Kent, W. J. (2012). The UCSC genome browser and associated tools. *Briefings in Bioinformatics*, *14*(2), 144–161.

Lee, C., Grasso, C., & Sharlow, M. F. (2002). Multiple sequence alignment using partial order graphs. *Bioinformatics* , *18*(3), 452–464.

Leonard, A. S., Crysnanto, D., Fang, Z.-H., Heaton, M. P., Vander Ley, B. L., Herrera, C., Bollwein, H., Bickhart, D. M., Kuhn, K. L., Smith, T. P. L., Rosen, B. D., & Pausch, H. (2022). Structural variant-based pangenome construction has low sensitivity to variability of haplotype-resolved bovine assemblies. *Nature Communications*, *13*(1), 1–13.

Liao, W.-W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J. K., Monlong, J., Abel, H. J., Buonaiuto, S., Chang, X. H., Cheng, H., Chu, J., Colonna, V., Eizenga, J. M., Feng, X., Fischer, C., Fulton, R. S., … Paten, B. (2022). A Draft Human Pangenome Reference. In *bioRxiv* (p. 2022.07.09.499321). https://doi.org/10.1101/2022.07.09.499321

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* , *34*(18), 3094–3100.

Li, H. (2019). Identifying centromeric satellites with dna-brnn. *Bioinformatics* , *35*(21), 4408–4410.

Li, H., Bloom, J. M., Farjoun, Y., Fleharty, M., Gauthier, L., Neale, B., & MacArthur, D. (2018). A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nature Methods*, *15*(8), 595–597.

Li, H., Feng, X., & Chu, C. (2020). The design and construction of reference pangenome graphs with minigraph. *Genome Biology*, *21*(1), 1–19.

Miga, K. H., & Wang, T. (2021). The Need for a Human Pangenome Reference Sequence. *Annual Review of Genomics and Human Genetics*, *22*, 81–102.

Miller, D. E., Kahsai, L., Buddika, K., Dixon, M. J., Kim, B. Y., Calvi, B. R., Sokol, N. S., Hawley, R. S., & Cook, K. R. (2020). Identification and Characterization of Breakpoints and Mutations on Drosophila melanogaster Balancer Chromosomes. *G3 Genes|Genomes|Genetics*, *10*(11), 4271–4285.

Mose, L. E., Wilkerson, M. D., Hayes, D. N., Perou, C. M., & Parker, J. S. (2014). ABRA: improved coding indel detection via assembly-based realignment. *Bioinformatics* , *30*(19), 2813–2815.

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., … Phillippy, A. M. (2022). The complete sequence of a human genome. *Science*, *376*(6588), 44–53.

Paten, B., Diekhans, M., Earl, D., St. John, J., Ma, J., Suh, B., & Haussler, D. (2011). Cactus Graphs for Genome Comparisons. *Journal of Computational Biology*, *18*(3), 461–489.

Paten, B., Earl, D., Nguyen, N., Diekhans, M., Zerbino, D., & Haussler, D. (2011). Cactus: Algorithms for genome multiple sequence alignment. *Genome Research*, *21*(9), 1512–1528.

Paten, B., Eizenga, J. M., Rosen, Y. M., Novak, A. M., Garrison, E., & Hickey, G. (2018). Superbubbles, Ultrabubbles, and Cacti. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, *25*(7), 649–663.

Paten, B., Herrero, J., Beal, K., Fitzgerald, S., & Birney, E. (2008). Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Research*, *18*(11), 1814–1828.

Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P. T., Gross, S. S., Dorfman, L., McLean, C. Y., & DePristo, M. A. (2018). A universal SNP and small-indel variant caller using deep neural networks.

Nature Biotechnology, *36*(10), 983–987.

Rautiainen, M., Nurk, S., Walenz, B. P., Logsdon, G. A., Porubsky, D., Rhie, A., Eichler, E. E., Phillippy, A. M., & Koren, S. (2022). Verkko: telomere-to-telomere assembly of diploid chromosomes. In *bioRxiv* (p. 2022.06.24.497523). https://doi.org/10.1101/2022.06.24.497523

Sirén, J., Garrison, E., Novak, A. M., Paten, B., & Durbin, R. (2020). Haplotype-aware graph indexes. *Bioinformatics* , *36*(2), 400–407.

Sirén, J., Monlong, J., Chang, X., Novak, A. M., Eizenga, J. M., Markello, C., Sibbesen, J. A., Hickey, G., Chang, P.-C., Carroll, A., Gupta, N., Gabriel, S., Blackwell, T. W., Ratan, A., Taylor, K. D., Rich, S. S., Rotter, J. I., Haussler, D., Garrison, E., & Paten, B. (2021). Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science*, *374*(6574), abg8871.

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

Vivian, J., Rao, A. A., Nothaft, F. A., Ketchum, C., Armstrong, J., Novak, A., Pfeil, J., Narkizian, J., Deran, A. D., Musselman-Brown, A., Schmidt, H., Amstutz, P., Craft, B., Goldman, M., Rosenbloom, K., Cline, M., O'Connor, B., Hanna, M., Birger, C., … Paten, B. (2017). Toil enables reproducible, open source, big biomedical data analyses. *Nature Biotechnology*, *35*(4), 314–316.

Wagner, J., Olson, N. D., Harris, L., McDaniel, J., Cheng, H., Fungtammasan, A., Hwang, Y.-C., Gupta, R., Wenger, A. M., Rowell, W. J., Khan, Z. M., Farek, J., Zhu, Y., Pisupati, A., Mahmoud, M., Xiao, C., Yoo, B., Sahraeian, S. M. E., Miller, D. E., … Sedlazeck, F. J. (2022). Curated variation benchmarks for challenging medically relevant autosomal genes. *Nature Biotechnology*, *40*(5), 672–680.

Zhou, Y., Zhang, Z., Bao, Z., Li, H., Lyu, Y., Zan, Y., Wu, Y., Cheng, L., Fang, Y., Wu, K., Zhang, J., Lyu, H., Lin, T., Gao, Q., Saha, S., Mueller, L., Fei, Z., Städler, T., Xu, S., … Huang, S.

(2022). Graph pangenome captures missing heritability and empowers tomato breeding. *Nature*, *606*(7914), 527–534.

Zook, J. M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., Weng, Z., Liu, Y., Mason, C. E., Alexander, N., Henaff, E., McIntyre, A. B. R., Chandramohan, D., Chen, F., Jaeger, E., Moshrefi, A., Pham, K., Stedman, W., Liang, T., … Salit, M. (2016). Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific Data*, *3*(1), 1–26.