bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

A Draft Human Pangenome Reference

Wen-Wei Liao[1,2,3,*], Mobin Asri[4,*], Jana Ebler[5,*], Daniel Doerr[5], Marina Haukness[4], Glenn Hickey[4], Shuangjia Lu[3], Julian K. Lucas[4], Jean Monlong[4], Haley J. Abel[6], Silvia Buonaiuto[7], Xian H. Chang[4], Haoyu Cheng[8,9], Justin Chu[8], Vincenza Colonna[7,10], Jordan M. Eizenga[4], Xiaowen Feng[8,9], Christian Fischer[10], Robert S. Fulton[1], Shilpa Garg[11], Cristian Groza[12], Andrea Guarracino[13], William T Harvey[14], Simon Heumos[15,16], Kerstin Howe[17], Miten Jain[18], Tsung-Yu Lu[19], Charles Markello[4], Fergal J. Martin[20], Matthew W. Mitchell[21], Katherine M. Munson[14], Moses Njagi Mwaniki[22], Adam M. Novak[4], Hugh E. Olsen[4], Trevor Pesout[4], David Porubsky[14], Pjotr Prins[10], Jonas A. Sibbesen[23], Chad Tomlinson[1], Flavia Villani[10], Mitchell R. Vollger[14,24], Human Pangenome Reference Consortium[#], Guillaume Bourque[25,26,27], Mark JP Chaisson[19], Paul Flicek[20], Adam M. Phillippy[28], Justin M. Zook[29], Evan E. Eichler[14,30], David Haussler[4,30], Erich D. Jarvis[31,30], Karen H. Miga[4], Ting Wang[32], Erik Garrison[10,+], Tobias Marschall[5,+], Ira Hall[3,33,+], Heng Li[8,9,+], Benedict Paten[4,+]

A complete list of authors and contributions appears at the end of the paper
* These authors contributed equally
+ Corresponding authors: egarris5@uthsc.edu, tobias.marschall@hhu.de, ira.hall@yale.edu, hli@jimmy.harvard.edu, bpaten@ucsc.edu
1 McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO 63108, USA
2 Department of Medicine, Washington University School of Medicine, St. Louis, MO 63110, USA
3 Department of Genetics, Yale University School of Medicine, New Haven, CT 06510, USA
4 UC Santa Cruz Genomics Institute, University of California, Santa Cruz, 1156 High St, Santa Cruz, CA, USA
5 Institute for Medical Biometry and Bioinformatics, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany
6 Division of Oncology, Department of Internal Medicine, Washington University School of Medicine, St. Louis, MO 63110, USA
7 Institute of Genetics and Biophysics, National Research Council, Naples 80111, Italy
8 Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA 02215, USA
9 Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02215, USA
10 Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, TN 38163, USA
11 Department of Biology, University of Copenhagen, Denmark
12 Quantitative Life Sciences, McGill University, Montreal, Québec H3A 0C7, Canada
13 Genomics Research Centre, Human Technopole, Milan 20157, Italy
14 Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA 98195, USA
15 Quantitative Biology Center (QBiC), University of Tübingen, Tübingen 72076, Germany
16 Biomedical Data Science, Department of Computer Science, University of Tübingen, Tübingen 72076, Germany
17 Tree of Life, Wellcome Sanger Institute, Hinxton, Cambridge, CB10 1SA, UK
18 Northeastern University, Boston, MA 02115, USA
19 University of Southern California, Quantitative and Computational Biology, Los Angeles, CA, USA

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

20 European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Cambridge, CB10 1SD, UK
21 Coriell Institute for Medical Research, Camden, NJ 08103, USA
22 Department of Computer Science, University of Pisa, Pisa 56127, Italy
23 Center for Health Data Science, University of Copenhagen, Denmark
24 Division of Medical Genetics, University of Washington School of Medicine, Seattle, WA 98195, USA
25 Department of Human Genetics, McGill University, Montreal, Québec H3A 0C7, Canada
26 Canadian Center for Computational Genomics, McGill University, Montreal, Québec H3A 0G1, Canada
27 Institute for the Advanced Study of Human Biology (WPI-ASHBi), Kyoto University, Kyoto 606-8501, Japan
28 Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA
29 Material Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, MD 20877, USA
30 Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA
31 The Rockefeller University, New York, NY 10065, USA
32 Department of Genetics, Washington University School of Medicine, St. Louis, MO 63110, USA

Abstract

The Human Pangenome Reference Consortium (HPRC) presents a first draft human pangenome reference. The pangenome contains 47 phased, diploid assemblies from a cohort of genetically diverse individuals. These assemblies cover more than 99% of the expected sequence and are more than 99% accurate at the structural and base-pair levels. Based on alignments of the assemblies, we generated a draft pangenome that captures known variants and haplotypes, reveals novel alleles at structurally complex loci, and adds 119 million base pairs of euchromatic polymorphic sequence and 1,529 gene duplications relative to the existing reference, GRCh38. Roughly 90 million of the additional base pairs derive from structural variation. Using our draft pangenome to analyze short-read data reduces errors when discovering small variants by 34% and boosts the detected structural variants per haplotype by 104% compared to GRCh38-based workflows, and by 34% compared to using previous diversity sets of genome assemblies.

Introduction

The human reference genome has formed the backbone of human genomics since its initial draft release more than twenty years ago (International Human Genome Sequencing Consortium, 2001). The primary sequences are a mosaic representation of individual haplotypes containing one representative scaffold sequence for each chromosome. There are 210 megabases (Mb) of gap/unknown (151 Mb) or computationally simulated sequence (59 Mb) within the current GRCh38 release, comprising 6.7% of the primary chromosome scaffolds (3.1 gigabases (Gb)). Missing reference sequences create an observational bias, or streetlamp

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

effect, which limits studies to be within the boundaries of the reference. Recently, the Telomere-to-Telomere (T2T) consortium finished the first complete sequence of a haploid human genome, T2T-CHM13, which provides a contiguous (no scaffold gaps) representation of each autosome and of chromosome X, with the exception of simulated ribosomal DNA arrays, totaling less than 10 Mb, that remain to be fully resolved (Nurk et al., 2022). Using T2T-CHM13 directly improves genomic analyses; for example, discovering 3.7 million additional single-nucleotide polymorphisms (SNPs) in regions non-syntenic to GRCh38 and better representing the true copy-number variants (CNVs) of 1000 Genomes Project (1KG) samples when compared to GRCh38 (1000 Genomes Project Consortium et al., 2015; Aganezov et al., 2022).

Although T2T-CHM13 represents a major achievement, no single complete genome can represent the genetic diversity of our species. Previous studies have identified tens of Mb of sequence contained within structural variants (SVs) that are polymorphic within the population (Ebert et al., 2021). Due to its repetitive nature and the absence of these alternative alleles and copy-number polymorphic paralogs from the reference genome, over two-thirds of SVs have been missed in studies using short-read data and the human reference assembly (Chaisson et al., 2019; Wenger et al., 2019; Zhao et al., 2021), despite individual SVs being more likely to impact gene function than either individual SNPs or short insertions and deletions (indels) (Chiang et al., 2017; Sudmant et al., 2015).

To overcome reference bias a transition to a pangenomic reference has been envisioned (Computational Pan-Genomics Consortium, 2018; Paten et al., 2017). Pangenomic methods have progressed rapidly over the last few years (Eizenga et al., 2020; Wang et al., 2022) such that it is now practical to propose migrating common genomic analyses to use a pangenome. Here, we sequence and assemble a set of diverse individual genomes and present a draft human pangenome, the first release of the Human Pangenome Reference Consortium (HPRC) (Wang et al., 2022). These genomes represent an initial subset of the planned HPRC panel, which aims to better capture global genomic diversity across the 700 haplotypes of 350 individuals.

Results

Assembling 47 diverse human genomes

We assembled 47 fully phased diploid assemblies from genomes selected to represent global genetic diversity (**Figure 1A**) and consented for unrestricted access data release. All are made publicly available, along with all data and analyses. The assembly process, as well as downstream quality control, were organized to ensure a high degree of completeness, contiguity, phasing, and base-level accuracy. These assemblies include 29 samples with long and linked read sequencing data generated entirely by the HPRC, and 19 samples sequenced by other efforts (Porubsky et al., 2021; Shafin et al., 2020; J. M. Zook et al., 2016), denoted HPRC+. In some cases we supplemented the HPRC+ samples with additional sequencing. We selected the 29 HPRC samples from the 1KG lymphoblastoid cell lines, limiting selection to

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

those lines classified as karyotypically normal, with low passage (to avoid artifacts from cell culture), and which were derived from participants for which whole-genome sequencing data is available for both parents (for phasing). Cell lines meeting these criteria were prioritized by genetic and biogeographic diversity (Methods).

We created a consistent set of deeply sequenced data types for every sample (**Supplementary Table 1**). The data included Pacific Biosciences (PacBio) high-fidelity (HiFi) and Oxford Nanopore Technologies (ONT) long-read sequencing, Bionano optical maps, and high coverage Hi-C Illumina short-read sequencing for all HPRC samples. We also gathered previously generated high-coverage Illumina sequencing data for both parents of each participant (Byrska-Bishop et al., 2021). We generated on average 39.7X HiFi sequence depth of coverage for the 46 HPRC/HPRC+ samples (excluding HG002, which had ~130X coverage) with a minimum of 30.6X (for HG02109) and maximum of 51.7X (for HG03453). This depth of coverage is consistent with the requirements for high-quality, state-of-the-art assemblies (Cheng et al., 2021) and facilitates comprehensive variant discovery irrespective of allele frequency. The N50 of the HiFi read lengths range from 13.5 kilobases (kb) to 26.9 kb (**Supplementary Table 1**), with an average of 19.6 kb (excluding HG002 because it was sequenced using a different library preparation protocol).

For the core assembler, we chose Trio-Hifiasm (Cheng et al., 2021) after detailed benchmarking of a large number of alternatives (Jarvis et al., 2022). Trio-Hifiasm uses PacBio HiFi long-reads and parental Illumina short-reads to produce near fully phased contig assemblies. The assembly process, as well as downstream quality control, were organized to ensure a high degree of completeness, contiguity, phasing, and base-level accuracy. The complete assembly pipeline (**Supplementary Figure 1**, Methods) includes steps to remove adaptor and non-human sequence contamination, and to ensure a single mitochondrial assembly per maternal assembly.

Assembly Assessment

We first searched for large-scale misassemblies, looking for large-scale gene duplication errors, phasing errors and interchromosomal misjoins (Methods). We fixed three large duplication errors and one large phasing error manually, while leaving smaller errors, which are hard to definitively distinguish from SVs. We found 217 putative interchromosomal joins. Only one of these joins (in the paternal assembly of HG02080) is located on euchromatic arms and was manually confirmed to be a misassembly. The remaining joins involve the short arms of the acrocentric chromosomes (**Figure 1B; Supplementary Table 2**) and may be the result of misalignment, nonallelic gene conversion, or other mechanisms that maintain large-scale homology between the short arms of the acrocentrics—a phenomenon which we study in detail in a companion manuscript (Guarracino, Buonaito, et al., 2022).

To evaluate the resulting assemblies after manual error correction, we developed an automated assembly quality control pipeline that combines methods to assess the completeness,

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

contiguity, base-level quality, and phasing accuracy of each assembly (Methods, **Supplementary Table 3**). Haploid assemblies containing an X chromosome have an average total length of 3.04 Gb, and match 99.3% of the length of the T2T-CHM13 assembly (3.06 Gb) that also contains an X chromosome. Haploid assemblies containing a Y chromosome average a total length of 2.93 Gb, reflecting the difference in size between the sex chromosomes (**Figure 1C**). The average NG50 – a widely used measure of contiguity, is 40 Mb – which is comparable to the 56 Mb NG50 of the contigs of GRCh38 (**Figure 1D**). Using short substrings (k-mers) derived from Illumina data, Yak (Cheng et al., 2021) estimates an average quality value (QV) of 53.57 for the assemblies, corresponding to an average of one base error per 227,509 bases. The assemblies vary in QV between 50 and 57, with a strong correlation between the maternal and paternal assemblies, as expected (**Figure 1E**). To help validate these QV estimates, we benchmarked the HG002 and HG005 assembly-based variant calls against the Genome in a Bottle (GIAB) v4.2.1 small variants; we estimated QVs as 54 for HG002 and 55 for HG005, similar to the k-mer QVs estimated by Yak of 53 and 54 for the HG002 maternal and paternal haplotypes, and 54 and 54 for the HG005 maternal and paternal haplotypes, respectively. Consistent with our manual observation that most errors were primarily small indels in low complexity regions, we found ~32% of indel errors were in homopolymers longer than 5 bp and an additional 48% were in tandem repeats and low complexity regions. Also, ~42% of indel errors were genotype errors, mostly heterozygous variants incorrectly called as homozygous variants due to collapsed haplotypes in the two assemblies of an individual (**Supplementary Table 4**). Analyzing the phasing accuracy between the maternal and paternal assemblies using k-mers derived from Illumina sequencing of the parents, Yak finds an average haplotype switch error rate of 0.67% and a Hamming error rate of 0.79% (**Figure 1F)**. We also calculated phase accuracy using Pstools (Garg, 2020, 2021), which uses Hi-C sequence data of the sample not used in creating the assembly. Pstools reports slightly lower switch error rates than Yak and comparable hamming error rates (**Supplementary Figure 2)**. Taken together, the above results indicate that the assemblies are highly contiguous and accurate.

Determining Regional Assembly Reliability

To determine which portions of the assemblies are reliable, we developed a read-based pipeline, Flagger, that detects different types of misassemblies within a phased diploid assembly (**Figure 1G**; Methods). The pipeline works by mapping the HiFi reads to the combined maternal and paternal assembly in a haplotype-aware manner and then identifying coverage inconsistencies within these read mappings that are likely due to assembly errors. This process is similar to likelihood-based approaches that assess the assembly given the reads (Rahman & Pachter, 2013), but is adapted to work with long-reads and a diploid assembly where both parental haplotypes are resolved. We identified only 0.88% (26.4Mb) of each assembly as unreliable based on Flagger analysis (**Figure 1H, Supplementary Table 5**). Compared to the distribution of contig sizes, the unreliable blocks are relatively short (54.6 kb N50 averaged across assemblies). The HG02572 paternal assembly contained the highest number of unreliable bases (~77.6 Mb) and the HG01175 maternal assembly contained the lowest number (~9.76 Mb). We intersected the Flagger unreliable blocks in the assemblies with different repeat annotations to measure the percentage of each annotation assembled confidently (**Figure 1I, Supplementary Table 6**). We estimate that 95.4% of alpha satellites, 91.5% of human satellites

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

2 and 3, 97.7% of segmental duplications (SDs), 94.3% of variable number tandem repeats (VNTRs), 94.2% of short tandem repeats (STRs), and 98.8% of all human repeats (Smit, AFA, Hubley, R & Green, P, 2013-2015) are correctly assembled. To estimate the false positive rate of Flagger, we ran Flagger on the CHM13 HiFi alignments to the T2T-CHM13 reference. Since this reference was extensively validated, we expected Flagger to report almost the whole assembly as reliable. After excluding rDNA arrays and satellites that are not properly evaluated using Flagger (Methods), but are also largely missing from the HPRC assemblies, we find 2.82 Mb (0.1%) of potentially (false positive) unreliable blocks in T2T-CHM13.

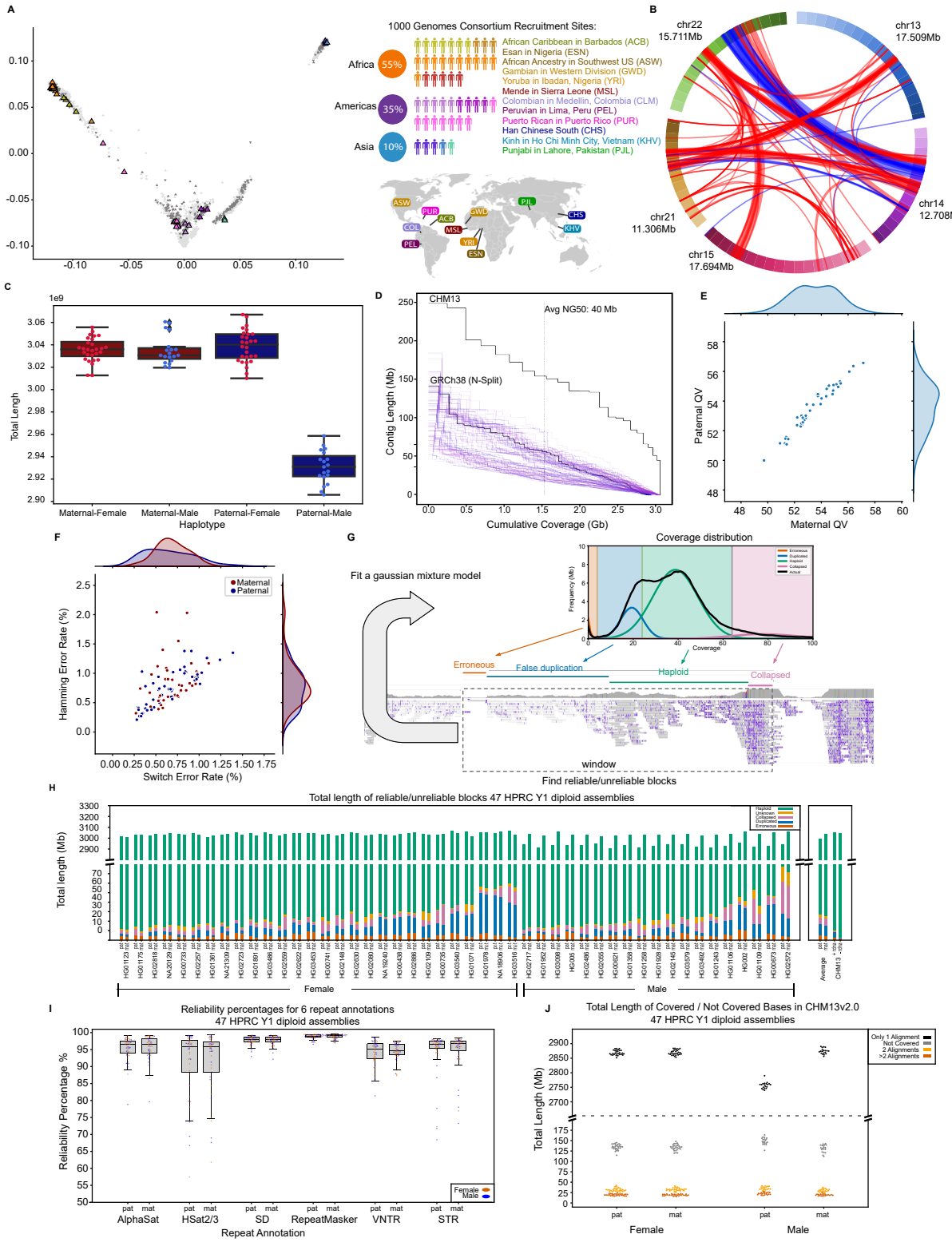
Assembly Completeness and Copy-Number Variation

To assess the completeness and copy-number polymorphism of the HPRC/HPRC+ assemblies we aligned them to the T2T-CHM13 reference and counted the number of reference bases with zero, one, two, and more than two alignments (Methods). The paternal assemblies of male samples cover ~92.8% of the reference (excluding chrX) on average with exactly one alignment; for all other assemblies (excluding chrY) ~94.1% on average is single-copy covered (**Figure 1J, Supplementary Table 7**). On average ~136 Mb (~4.4%) of the T2T-CHM13 reference is not covered by any alignment, showing that some parts of the genome are either systematically unassembled or cannot be reliably aligned to; ~90% of these regions are peri/centromeric (**Supplementary Figure 3**), with the active/inactive alpha satellites and human satellite 3 comprising about ~50% of these bases, mainly due to their highly repetitive composition and also higher frequency compared to other satellites (Altemose et al., 2022). Other centromeric satellites, centromeric transition regions, and rDNA arrays accounted for another ~40% of the uncovered bases on average. Despite the majority of unaligned bases occurring within and around centromeres, on average 90% of divergent/monomeric alpha satellites, gamma satellites, and centromeric transition regions are covered by at least one alignment. On the other hand, rDNA arrays, which are by far the hardest repeat arrays to assemble, were the least covered repeat array (~8%). Excluding the T2T-CHM13 centromere and satellites (Nurk et al., 2022) and including only the expected sex chromosome for each haploid assembly, on average ~99.12% of the remaining reference is covered by exactly one alignment **(Supplementary Table 7)**.

The average number of the T2T-CHM13 bases with two and more than two alignments are ~32.4 Mb (~1.0%) and ~20.0 Mb (~0.6%), respectively. On average per haploid assembly, these duplicated alignments had ~82.20% and ~39.82% overlap with the peri/centromeric satellites and SDs, respectively, and ~94.62% had overlap with either of them. Many of these duplicated alignments correspond to SDs. We then characterized the accuracy of regions aligned to SDs in T2T-CHM13 (excluding chromosome Y) using a liftover of the assembly read-depth based evaluation (**Supplementary Figure 4)**. On average we estimate that only 2.5% (4.99/199 Mb) of the SD sequence that can be lifted onto T2T-CHM13 is in error according to read depth. To identify SDs associated with these errors, we took all 5 kb windows across the unreliable regions and intersected them with the longest and most identical overlapping SD. The median length of SDs overlapping sequences in error is 3.0 times longer (288 kb vs. 96.3 kb) than those in correctly assembled SDs and 1.8% more identical (98.9 vs. 97.1), reinforcing

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

earlier findings that length and identity of SDs play an important role in assembly accuracy (Alkan et al., 2011).



bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 1 | 47 accurate and near-complete diverse diploid human genome assemblies. A) Selecting the HPRC samples. Left: The first two principal components of 1KG samples showing HPRC/HPRC+ (triangles) samples. Right: A summary of the HPRC/HPRC+ samples and their subpopulations (three letter abbreviations on Earth map) as defined by 1KG. **B)** Interchromosomal joins between acrocentric chromosome short arms. Red means the join is on the same strand; blue otherwise. **C)** Total assembled sequence per haploid phased assembly. **D)** Assembly contiguity shown as an NGx plot. T2T-CHM13 and the contigs of GRCh38 are included for comparison. **E)** Assembly quality values (QV), showing the base-level accuracy of each sample's maternal and paternal assembly. **F)** Yak reported phasing accuracy**,** showing switch error % vs. Hamming error %. **G)** An overview of the Flagger read-based assembly evaluation pipeline. Coverage is calculated across the genome and a mixture model is fit to account for reliably assembled haploid sequence as well as various classes of unreliably assembled sequence. For each coverage block a label is assigned according to the most probable mixture component it belongs to: erroneous, (falsely) duplicated, (reliable) haploid, collapsed, and unknown. **H)** Estimating the reliability of the 47 HPRC/HPRC+ assemblies using read mapping. Regions flagged as haploid are reliable (colored green), on average they constitute more than 99% of each assembly. The Y axis is broken to show the dominance of the reliable haploid component and also the stratification of the unreliable blocks. **I)** Evaluating the assembly reliability of 6 repeat annotations. The regions flagged as haploid by Flagger were intersected with 6 repeat annotations; the repeat annotations included alpha satellites, human satellites 2 and 3, SDs, the repeats characterized by RepeatMasker (Smit, AFA, Hubley, R & Green, P, 2013-2015), VNTRs, and STRs. **J)** Completeness of the HPRC assemblies. We aligned all HPRC assemblies to the T2T-CHM13 reference to assess their completeness. The number of reference bases covered by no, one, two, and more than two alignments are counted separately.

Annotating 47 Diverse Genome Assemblies

We developed a new Ensembl mapping pipeline to annotate GENCODE (Frankish et al., 2021) genes and transcripts within each new haploid assembly (Methods). To create high-confidence annotations, the pipeline clusters and maps spatially proximal genes in parallel (to help avoid issues with individually mapping near identical paralogues) and attempts to resolve inconsistent mappings by both considering the synteny of the gene neighborhood in relation to the GRCh38 annotation and the identity and coverage of the underlying mappings. A median of 99.07% of protein-coding genes (minimum of 98.08%, maximum of 99.40%) and 99.42% of protein-coding transcripts (minimum of 98.29%, maximum of 99.66%) were unambiguously identifiable in each of the HPRC assemblies (**Figure 2A; Supplementary Table 8**). Similarly, a median of 98.16% of non-coding genes (minimum of 97.23%, maximum of 98.60%) and 98.96% of non-coding transcripts (minimum of 97.94%, maximum of 99.28%) were similarly annotated. By way of comparison, running this pipeline on T2T-CHM13 gives comparable, if slightly higher, results: we annotated 99.54% and 99.76% of protein-coding genes and transcripts, and 99.11% and 99.52% of non-coding genes and transcripts in T2T-CHM13. Intersecting the HPRC/HPRC+ annotations with the assembly reliability predictions, a median of 99.53% of gene and 99.79% of transcript annotations occurred wholly within reliable regions, indicating that the vast majority of the annotated haplotypes are structurally correct. To examine transcriptome base accuracy, we looked for nonsense and frameshift mutations in the set of canonical transcripts (one

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

representative transcript per gene; Methods and **Supplementary Table 8, Supplementary Figure 5**). We found a median of 25 nonsense mutations per assembly, supporting the idea that there is a low level of base-level substitution error. A median of 21 (84%) of these nonsense mutations per assembly are supported by the independently generated Illumina variant call sets. We found a median of 72 frameshifts (0.37% of transcripts) mutations per genome, with a median of 67 of these being high-confidence frameshifts not occurring in the leading 5' or 3' ends of the transcript. A median of 58 (80%) of these frameshifts per assembly are supported by the same Illumina call sets. These numbers are within the range of previously reported numbers of loss-of-function mutations (between 10-150 per person, depending on the level of conservation of the mutation) (1000 Genomes Project Consortium et al., 2015; MacArthur et al., 2012). Some of the non-confirmed frameshifts and nonsense mutations (a median of 14 frameshifts and 4 nonsense mutations per assembly, or one error per ~1.7 million reference transcriptome bases) are likely assembly errors.

There are 1,529 protein-coding gene families within the Flagger predicted reliable regions of the full set of assemblies that have a gain in copy number in at least one genome (**Figure 2B**). Each assembly has an average of 44 genes with a gain in copy number relative to GRCh38 within its predicted reliable regions, with a bias towards rare, low-copy CNVs (**Figure 2C**); 80% of CNV genes appear in a single haplotype. Previous studies using read depth found that rare CNVs occur generally outside of regions annotated as being enriched in SDs (Sudmant et al., 2010). The genome assemblies confirm this observation in sequence-resolved CNVs. When stratifying duplicated genes based on allele frequency (AF) into singleton (present in one haplotype), low frequency (< 10%), and high frequency, 13% (159/1,181) of the singleton CNVs map to SDs as annotated in GRCh38. Duplicated genes with a higher population frequency have a greater fraction in SDs: 40% (86/214) of low-frequency, and 81% (148/184) of high frequency. 63 genes are CNVs in 10% or more of haploid assemblies, and 17 genes are amplified in the majority of individuals relative to GRCh38 (**Figure 2D; Supplementary Table 9**). Many of these genes are individually highly copy-number polymorphic and part of complex tandem duplications (**Figure 2E**). For example, the gene *GPRIN2* is known to be copy-number polymorphic (Handsaker et al., 2015) based on read depth, and has sequence resolution of 1-3 additional copies duplicated in tandem in the pangenome (**Figure 2F)**. The gene *SPDYE2* is similarly resolved as 1-4 additional copies duplicated in tandem (**Figure 2G**). Other copy number variable genes are not contiguously resolved and reflect limitations of the current assemblies (see Porubsky et al. companion). For example, the defensin gene *DEFB107A* has 3-8 additional copies assembled across all samples, however this gene is assembled into 3-7 separate contigs that do not reflect the global organization of this gene.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

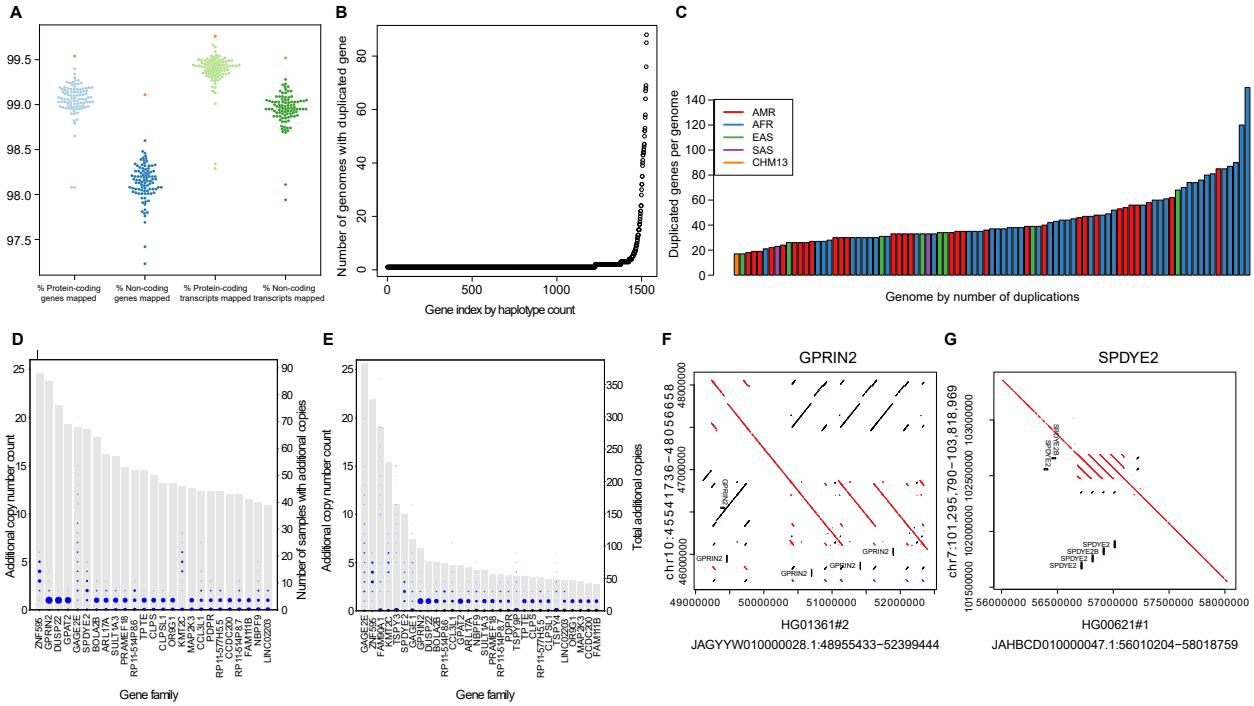


Figure 2 | Transcriptome annotation of the assemblies. A) Ensembl mapping pipeline results. Percentages of protein-coding and non-coding genes and transcripts annotated from the reference set in each of the HPRC assemblies. Orange points represent T2T-CHM13 for comparison. **B)** Assembled gene duplications per genome. The number of genomes containing a duplicated gene for 1529 protein-coding gene duplications indexed by increasing copy number, observed in the predicted reliable regions of the HPRC/HPRC+ genomes. **C)** Number of distinct duplicated genes or gene families per phased assembly relative to the number of duplicated genes annotated in GRCh38 (152). The GRCh38 gene duplications reflect families of duplicated genes, while the counts in other genomes reflect gene duplication polymorphisms. The assemblies are color coded according to their population of origin. **D)** The top 25 most commonly CNV genes or gene-families in the HPRC/HPRC+ assemblies, ordered by the number of samples with additional copies relative to GRCh38. Grey bars represent the number of samples with additional copies. Blue circles represent the number of additional copies per sample, with the size of the circle proportional to the number of samples. **E)** The top 30 most individually CNV genes or gene families in the HPRC/HPRC+ assemblies, ordered by total number of additional copies observed. Blue circles again represent the number of additional copies per sample, with the size of the circle proportional to the number of samples. The Grey bars represent the total number of additional copies summed over the samples. **F)** Dotplot illustrating haplotype-resolved GPRIN2 gains in the HG01361 assembly relative to GRCh38. **G)** Dotplot illustrating SPDYE2/SPDYE2B haplotype resolved gains within a tandem duplication cluster of the HG00621 assembly relative to GRCh38.

Constructing a draft pangenome

We use a sequence graph representation for pangenomes (Eizenga et al., 2020; Paten et al., 2017) in which nodes correspond to segments of DNA. Each node has two possible orientations, forward and reverse, and there are four possible edges between any pair of nodes to reflect all combinations of orientations (bidirected graph). The underlying haplotype sequences can be represented as walks in the graph. The model represents a generalized

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

multiple alignment of the genome assemblies from which we build it: haplotypes are aligned where they co-occur on a given node (**Figure 3A**).

The process of generating a combined pangenome representation is non-trivial because determining which alignments to include is not always obvious, particularly for recently duplicated and repetitive sequences. We applied three different graph construction methods: Minigraph (Li et al., 2020), Minigraph-Cactus (MC), and PanGenome Graph Builder (PGGB) (Methods). The availability of these three models provides us with multiple views into the homology relationships in the pangenome while supporting cross-validation of discovered variation. We included the GRCh38 and T2T-CHM13 assemblies within the pangenomes and three samples were held out from the pangenome graphs to permit their use in benchmarking: HG002, HG005, and NA19240.

Minigraph and Minigraph-Cactus

Minigraph builds a pangenome by starting from a reference assembly, here GRCh38, and iteratively and progressively adds in additional assemblies, recording only SVs larger than or equal to 50 bases. It admits complex variants, including duplications and inversions. MC extends the Minigraph pangenome with a base-level alignment of the homology relationships between the assemblies using the Cactus genome aligner (Armstrong et al., 2020), while retaining the structure of the Minigraph pangenome. To remove noisy alignments from the MC pangenome, long (≥100 kb) non-reference sequences identified either as being satellite, unassignable to a reference chromosome, or which appear unaligned to the remainder of the assemblies, were removed from the graph. The result is a pangenome with significantly reduced complexity that nevertheless maintains all sequences of the starting reference assembly and the large majority of those in the additional haplotypes.

The PanGenome Graph Builder

PGGB constructs a pangenome from an all-to-all alignment of the assemblies. Subsequent stages of refinement compress and normalize the graph using partial order alignment (Gao et al., 2021). Although both the CHM13 and GRCh38 reference are used to partition contigs into chromosomes, the PGGB graph does not base itself on a chosen reference assembly, and includes both references and all HPRC haplotypes in a single graph. Due to ambiguous placement of variation in all-to-all pairwise alignments, many SV hotspots, including the centromeres, are transitively collapsed into loops through a subgraph representing a single repeat copy, a feature which tends to reduce the size of variants found in repetitive sequences. The PGGB graph provides a lossless representation of the input assemblies, without filtering of rapidly-evolving satellite sequences or clipping of regions that do not reliably align. This increases its size and complexity relative to the MC graph, and adds a significant amount of “singleton” sequence relative to the Minigraph and MC graphs. However, this property allows for annotations and coordinates of all assemblies in the pangenome to be related to the graph structure and utilized in subsequent downstream analyses.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Measuring Pangenome Variation

The different algorithmic approaches for constructing a pangenome graph influence graph properties while representing the same underlying sequences. The basic properties of the three graphs produced with the different pangenome methods are shown in **Supplementary Table 10**. The Minigraph graph, by virtue of being limited to structural variation, is smallest, with over two orders of magnitude fewer nodes and edges than the base-level graphs. Its length (3.24 Gb), measured as the total bases of all nodes, is similar to the MC graph (3.29 Gb) because while the latter adds many small variants it also aligns together a significant number of sequences that were left unaligned by Minigraph. The PGGB graph contains roughly five gigabases more sequence because it includes highly structurally divergent satellite regions omitted from the other approaches, and does not implement any trimming or filtering of the input assembly contigs.

To characterize the variants in these pangenome graphs we used graph decomposition to identify “bubble” subgraphs that correspond to non-overlapping variant sites (Methods). In the MC and PGGB graphs we classified the variant sites into small variants (<50 bp) and SVs (≥50 bp); SV sites in all three graphs were further classified into various repeat classes using the longest allele for each site. We found similar numbers of each variant type in each pangenome, 22 (21) million small variants in the MC (PGGB) graphs (**Figure 3B**), and 67 (73, 75) thousand SVs in the MC (PGGB, Minigraph) graphs (**Figure 3C**). We find a total of 90 (55, 86) megabases of non-reference sequence in the SV sites, excluding centromeric repeats (which are difficult to align), in the MC (PGGB, Minigraph) graph. Clustering the SV alleles by length and similarity, Alu, L1 and ERV SVs appear largely biallelic, however VNTRs frequently have three or more distinct alleles per site. The minor AF in the pangenomes of biallelic variants is similar for SNPs as well as L1s, Alus and VNTRs, although VNTRs show a slight shift toward more common alleles (**Figure 3D**).

The MC and PGGB pangenome graphs encode the underlying 44 diploid genome assemblies used in their construction as paths within the graph. For these pangenomes it is therefore possible to trace each of these assemblies within the graph and decode their alleles as they visit variant sites (Methods). We find similar numbers of small variants and SVs in the Dipcall confident regions: 5.34 (5.35) million small variants per sample and 16.8 (17.4) thousand SVs per haplotype on average in the MC (PGGB) graph (**Figure 3E-F**), with the differences in the numbers of such variants recapitulating previously observed differences between the samples that are a result of their ancestry (1000 Genomes Project Consortium et al., 2015).

We quantified the amount of euchromatic autosomal non-reference (GRCh38) sequence that each of the 44 diploid genomes incrementally contributes to the pangenome (**Figure 3G**, Methods) for both graphs. We limit to the euchromatic sequence because we are generally confident in its assembly and alignment. We roughly approximate the euchromatic regions as the sequences included in the MC graph, since the MC graph omits common centromeric satellites and other sequences that failed to align. Since the PGGB graph includes all assembled sequences for this analysis we pruned regions not contained in the MC graph from the PGGB graph. We then grouped genomes by their assigned super populations and

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

computed cumulative base pair lengths from the 1st to the 44th genome. Overall, the euchromatic autosomal non-reference sequence adds up to ~175 Mb in MC (and ~190 Mb in PGGB), out of which ~55 Mb (~105 Mb) are observed only on a single haplotype. Our analysis further suggests that ~5 Mb and ~70 Mb (~10 Mb and ~60 Mb) can be attributed to core (present in ≥95% of all haplotypes) and common genome (present in ≥5% of all haplotypes), respectively (**Supplementary Table 11**). We additionally estimate the growth of the euchromatic autosomal pangenome independent of genomes' order. To this end, we sampled 200 permutations of genome orderings (**Supplementary Figure 6**) and recorded the median pangenome size across all samples in the MC graph. Our results indicate that the second genome adds ~23 Mb of euchromatic autosomal sequence to the pangenome while the last genome tends to add much less with only about ~0.64 Mb. These numbers are conservative, owing to additional highly polymorphic sequence residing in the sequence gaps of our assemblies. Extrapolating under Heaps' Law ("Comparative Genomics: The Bacterial Pan-Genome," 2008) (Methods), we expect at least an additional ~150 Mb of sequence in the pangenome graph when HPRC produces 700 haplotypes in future.

We annotated the small variants overlapping the GIAB v3.0 "easy" regions (covering 74.35% of the primary chromosome scaffolds of GRCh38) with AFs from gnomAD v3.1.2 (**Figure 3H; Supplementary Table 12**). These variants are generally straightforward to annotate accurately. In the MC graph, about 60.2% (~9.7 million variants) have an AF of 1% or greater. About 35.7% are rare, having an AF less than 1% but above zero. About 1.7% are singleton. The remaining 2.4% are missing from gnomAD. We find similar results with the PGGB graph, repeating this exercise with small variant calls by pairwise alignment of the assemblies directly to the reference using Dipcall (Li et al., 2018), and by calling small variants from the HiFi sequencing data using DeepVariant (Poplin et al., 2018). These missing variants are therefore likely mostly a mixture of variants missing from gnomAD and assembly errors.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

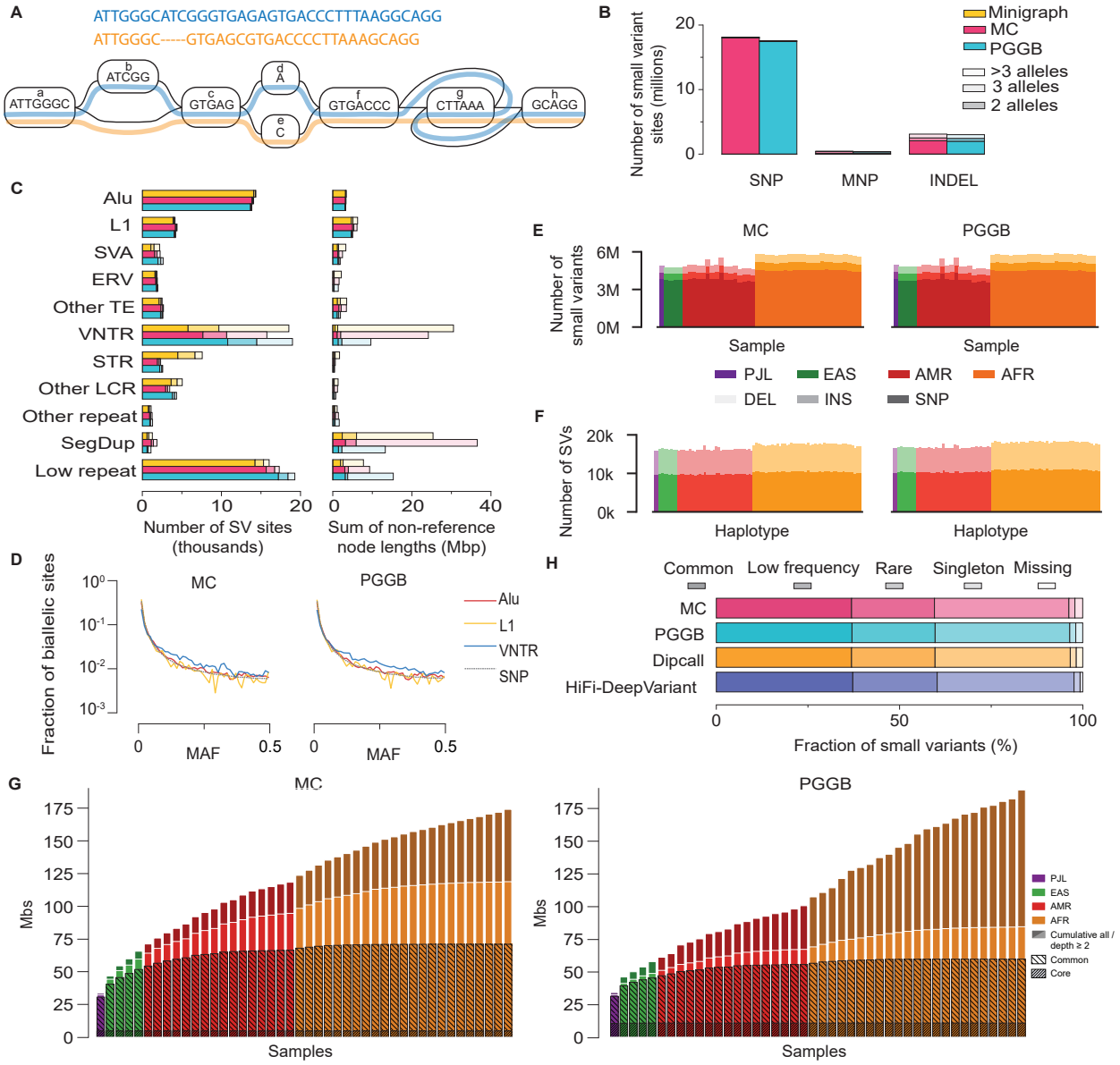


Figure 3 | Pangenome Graphs Represent Diverse Variation. A) A pangenome variation graph. It is comprised of two elements: a sequence graph, whose nodes represent oriented DNA strings and whose bidirected edges represent the connectivity relationships, and embedded haplotype paths (colored lines) that represent the individual assemblies. **B)** Small variant sites in pangenome graphs, stratified by the variant type and by the number of alleles at each site. MNP: multi-nucleotide polymorphism. **C)** SV sites in pangenome graphs stratified by repeat class and by the number of alleles at each site. Other TE: a site involving mixed classes of transposable elements. VNTR: variable-number tandem repeat, a tandem repeat with the unit motif length ≥7bp. STR: short tandem repeat, a tandem repeat with the unit motif length ≤6bp. Other LCR: low-complexity regions with mixed VNTR/STR and low-complexity regions without a clear VNTR/STR pattern. Other repeat: a site involving mixed classes of repeats. SegDup: segmental duplication. Low repeat: a small fraction of the longest allele in a site involving repeats. **D)** Pangenome minor AF (MAF) spectrum for bi-allelic SNPs, VNTRs, L1s and Alus in the MC and PGGB graphs. **E-F)** Number of autosomal (E) small variants per sample and (F) SVs per haplotype in the pangenome. Variants restricted to the Dipcall confident regions.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Samples organized by 1KG populations. **G)** Pangenome growth curves for MC (left) and PGGB (right). Depth measures how often a segment is contained in any haplotype sequence, core is present in ≥95% of haplotypes, common is ≥5%. **H)** Small variants in the GIAB v3.0 easy regions annotated with allele frequencies from gnomAD v3.1.2.

To further explore the quality of variant calls captured by assembly and graph construction, we compared pangenome decoded variants against GRCh38 to variant sets identified by conventional reference-based genotyping methods (**Supplementary Figure 7**, Methods). These reference-based call sets were generated from the PacBio HiFi reads and haplotype-resolved assemblies using different discovery methods: DeepVariant (Poplin et al., 2018), PBSV (Pacific Biosciences, 2021), Sniffles (Sedlazeck et al., 2018) with Iris (Kirsche et al., 2021), SVIM (Heller & Vingron, 2019), SVIM-asm (Heller & Vingron, 2020), PAV (Ebert et al., 2021), and the Hall-lab pipeline (Methods). For benchmarking small variants, we excluded regions containing SVs detected or implied by the alignment of that sample's Hifiasm assembly to GRCh38, since current benchmarking tools do not account for different representations of small variants inside or near SVs (Methods). Comparing small variants (**Figure 4A**) and SVs (**Figure 4B**) from the pangenomes to the reference-based sets we see a high level of concordance that varies, as expected, by the relative repeat content of the surrounding genome. Overall, variant calling performance is extremely high in both the MC and PGGB graphs. For example, in relatively unique 'easy' genomic regions comprising 75.42% of the autosomal genome, samples show a mean of 99.64% recall and 99.64% precision for small variants in the MC graph, and in 'high confidence' regions (~90% of autosomal genome) they show 97.91% and 96.66%, respectively (**Figure 4A**). Performance is somewhat lower for SVs than for small variants (**Figure 4B**), as expected, but is still strong. Variant calling performance diminishes substantially (but is still respectable) in highly repetitive genome regions (3.87% of autosomal genome) (**Figure 4A-B**), for which more work will be required to achieve high quality variant maps. We further note that these are likely to be significant underestimates of variant calling quality considering known errors in the truth set due to the inherent limitations of reference-based variant callers (see below). Stratifying the insertion and deletion variants within the pangenome we observe relatively constant, high levels of agreement with the reference-based methods regardless of length (**Figure 4C).**

An independent measure of pangenome graph quality is the extent to which sample haplotype paths through the graph are well supported by the raw sequencing data. We calculated the number of supporting reads by aligning them to the MC graph using GraphAligner (Methods; the inclusion of heterochromatic sequences in the PGGB graph made read mapping impractically slow). We found that over 97% of HiFi reads can be aligned to the MC graph after filtering (**Figure 4D**, left). Among these aligned reads, we further calculated the read depth of on- and off-target edges based on the sample paths in the graph. We found that on average over 94% of on-target edges were supported by at least five reads and observed two peaks in the read depth distribution of on-target edges (**Figure 4D**, middle): a minor peak corresponding to the edges in heterozygous regions and a major peak at twice the minor peak corresponding to the edges in homozygous regions. In contrast, only 7% or fewer off-target edges were supported by at least five reads (**Figure 4D**, right). In addition to HiFi reads, we also used ONT reads from 29 out of

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the 44 samples to perform the same analysis and, despite the data being lower coverage, found similar results (**Supplementary Figures 8 and 9**).

These data also show that the pangenome graphs perform better at capturing genome variation than the above benchmarking results imply. For example, a mean of 89.3% of putative false positive small variant calls are supported by ≥5 HiFi reads, and 75.3% by ≥10 reads (85.9% and 73.8% for SVs), suggesting that most putative errors are in fact real variants that were missed by the reference-based callers used to create the truth set (**Supplementary Figure 10**; **Supplementary Table 13**).

We used the Comparative Annotation Toolkit (CAT) (Fiddes, Armstrong, et al., 2018) to lift-over GENCODE v38 annotations using the MC pangenome graph onto the individual haplotype assemblies. CAT lifted and annotated a median of 99.5% of 86,757 protein-coding transcripts per assembly (Methods**, Supplementary Figures 11 and 12, Supplementary Table 14)**, almost the same as the Ensembl mapping based pipeline (a median of 99.4% per assembly), supporting the idea that the MC pangenome captures most transcript homologies.

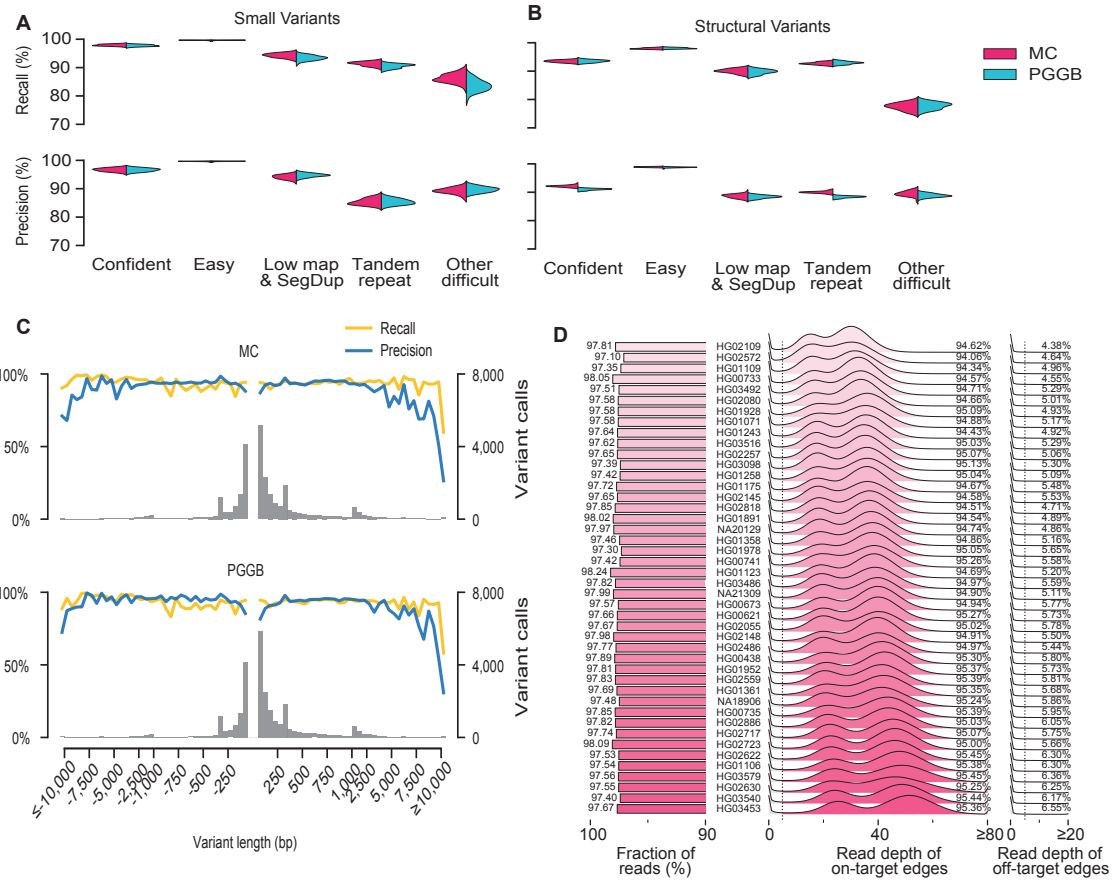


Figure 4 | Pangenome graph evaluation. A-B) The precision and recall of autosomal (**A**) small variants and (**B**) SVs in the pangenomes relative to consensus variant sets. Small variants are compared to HiFi-DeepVariant calls. SVs are compared to the consensus of six reference-

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

based SV callers (Methods). Comparisons are restricted to the Dipcall confident regions and then stratified by the GIAB v3.0 genomic context. **C)** Average SV precision, recall, and frequency in the Dipcall confident regions stratified by length, in the MC (top) and PGGB (bottom) graph relative to consensus SV sets. The histogram bin size is 50 bp for SVs <1 kb, and 500 bp for SVs ≥1 kb. **D)** HiFi read depth of on- and off-target edges in the MC graph. Left: fraction of reads aligned to the pangenome graph after filtering low-quality alignments. Middle: read depth distribution of on-target edges. Right: read depth distribution of off-target edges. Samples are sorted by sequencing coverage (**Supplementary Table 1**).

Pangenomes Represent Complex Loci

We annotated and visualized the structure of haplotypes in 5 clinically relevant multiallelic CNV loci, RHD/RHCE, HLA-A, CYP2D6/CYP2D7, C4, and LPA within the PGGB and MC pangenomes (Methods). For each locus and graph, we identified its location within the graph and then annotated paths within this subgraph with known genes. We then traced the individual haplotypes through the subgraph to reveal the structural haplotypes of each assembly. In CYP2D6/7 (**Supplementary Figure 13**), C4 (**Supplementary Figure 14**), and LPA (**Supplementary Figure 15**), we recapitulated previously described haplotypes. For CYP2D6/7, our calls matched 96% of haplotypes of 76 assemblies called by Cyrius using Illumina short-reads data (Chen et al., 2021). Two discrepancies appear to be caused by errors from Cyrius, and the third is a false duplication in the HG01071#2 pangenome assembly revealed by Flagger. This comparison suggests the pangenomes faithfully agree with existing knowledge of this complex loci. In RHD/RHCE (**Figure 5, top**), in addition to previously described haplotypes, we also inferred the presence of 5 novel haplotypes, which included one duplication allele of the RHD gene, and one inversion allele occurring between the RHD and RHCE gene that results in the swapping of the last exon of both genes. Around HLA-A (**Figure 5, bottom; Supplementary Figure 16**), two deletion alleles have been previously described – albeit with imprecise breakpoints (Sudmant et al., 2015) – but an insertion allele carrying an HLA-Y pseudogene is previously unreported. The long sequence (65 kb) inserted with HLA-Y occurs at high frequency (28%) but has little homology to GRCh38. We compared the representation of these 5 loci in the MC and PGGB graphs (**Supplementary Figure 17**). Each graph independently recapitulated the same haplotype structures, but in general the PGGB graphs tend to use a single collapsed copy to represent adjacent homologous sequences. Assemblies that contain multiple copies of the homologous sequence traverse these nodes a corresponding number of times. MC maintains separate copies of these homologous sequences.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

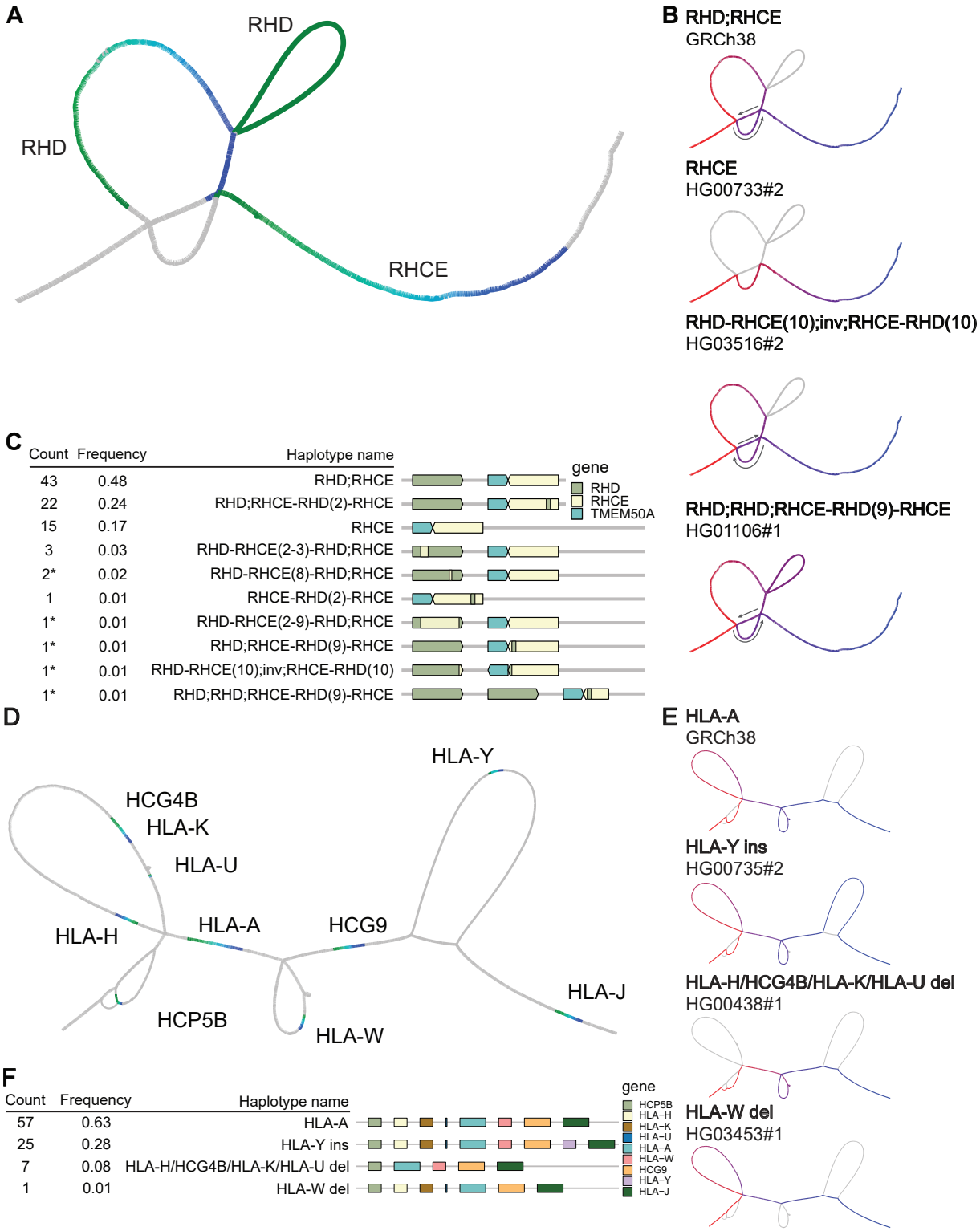


Figure 5 | Visualizing Complex Pangenome Loci. At top are structural haplotypes of RHD/RHCE called from the MC graph. **A)** Location of the RHD and RHCE gene within the MC subgraph. The color gradient is based on the relative position of a gene. Green represents the head of a gene. Blue represents the end of a gene. **B)** Different structural haplotypes take different paths through the graph. The color gradient is based on path position; red represents

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the start of a path; blue represents the end of a path. **C)** Frequency and linear structural visualization of all structural haplotypes called by MC graph among 90 haploid assemblies. Asterisks in the count column indicate the novel haplotypes we found. Shown at bottom are structural haplotypes of HLA-A called from the PGGB graph. **D)** Location of genes within the PGGB subgraph. **E)** Different structural haplotypes take different paths through the graph. **F)** Frequency and linear structural visualization of all structural haplotypes called by the PGGB graph.

Applications of the pangenome

Pangenome-based short variant discovery

Our pangenome reference aims to broadly improve downstream analysis workflows by removing mapping biases inherent to the use of a single linear reference genome such as GRCh38 or CHM13. As a first use case, we studied whether mapping against our pangenomes could improve small variant calling accuracy from short-reads. We aligned short-reads from the GIAB benchmark samples (J. M. Zook et al., 2016) to the MC pangenome graph with vg Giraffe (Sirén et al., 2021). For comparison, we aligned reads to GRCh38 using BWA-MEM (Li, 2013) and to Dragen Graph (Miller et al., 2015), which uses GRCh38 augmented with alternative haplotypes at variant sites. We called SNPs and indels with DeepVariant (Poplin et al., 2018) and the Dragen variant caller (Miller et al., 2015) (Methods). Our pangenomic approach (Giraffe+DeepVariant) outperforms the other approaches on small variants (**Figure 6A**), with gains for both SNPs and indels (**Supplementary Figure 18, Supplementary Table 15**). For example, it made only 21,700 errors (false positives or false negatives) in the confident regions of the GIAB truth set using 30x reads from HG005. In contrast, 36,144 errors were made when DeepVariant used the reads aligned to GRCh38, and 26,852 errors when using the Dragen pipeline. In challenging medically relevant genes (Wagner, Olson, Harris, McDaniel, et al., 2022), the increase in performance is even larger for both SNPs (F1 score of 0.985 for Giraffe-DeepVariant vs <0.976 for other methods) and indels (F1 score of 0.961 for Giraffe-DeepVariant vs <0.958 for other methods) (**Figure 6B**). Many regions benefit from using pangenome mapping, but regions with errors in GRCh38 and large L1HS sequences benefit the most from the pangenomic approach (**Supplementary Figure 19**). Although the genotyping performance was lower for the short variants called that were not present in the pangenome, the pangenomic approach suffered less than other approaches, suggesting that novel variants also benefit from the pangenome (**Supplementary Figure 20**).

We next benchmarked variant calling using parent-child trios. Using DeepTrio (Kolesnikov et al., 2021) resulted in better performance relative to DeepVariant across all samples of the GIAB (**Figure 6A**) and the challenging medically-relevant genes benchmarks (**Figure 6B**, **Supplementary Figure 18)**, with improvements that appear additive to those from the pangenome. For example, DeepTrio using HPRC+Giraffe alignments gave the highest calling accuracy, with the number of errors decreasing from 21,700 (single sample calling) to 10,098 (trio calling) for HG005.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

A pangenome resource across the 1000 Genomes Project cohort

To create a community resource aiding method development and pangenome-based population genetic analyses, we used Giraffe to align high-coverage short-read data from 3,202 samples of the 1KG (Byrska-Bishop et al., 2021) to our pangenome graph and DeepVariant to call small variants (Methods/Data Availability). The Mendelian consistency computed across 100 trios from those samples was comparable to the one computed across samples from the GIAB truth set, indicating a comparable call set quality (**Supplementary Figure 21 and 22**). Given that our pangenome-based calls showed superior performance in challenging regions (**Figure 6B**), this call set across the 1KG cohort now provides the genetics and genomics communities with AF estimates for complex but medically-relevant loci. For instance, our approach was able to detect the gene conversion event covering the second exon of the *RHCE* gene, which was observed in about 25% of assembled haplotypes (**Figure 5C; Supplementary Figure 23**). The gene-converted allele is indeed in the pangenome (**Supplementary Figure 23 d-e**) and a cluster of variants with 25% frequency is now visible (**Supplementary Figure 23 c**). Also in the *KCNE1* gene, we provide calls and frequencies in a 40 kb region, spanning three exons, that could not be assessed before due to the presence of a false duplication in GRCh38 (**Supplementary Figure 24**; See Vollger *et al.* companion for genome-wide analysis of interlocus gene conversion).

Pangenome-based structural variant genotyping

The ability to represent polymorphic SVs is a key advantage of a graph-based pangenome reference. To demonstrate the utility of the sequence-resolved SVs inherent to our pangenome, we used PanGenie (Ebler et al., 2022) to genotype the bubbles in the MC graph. We decomposed bubbles into their constituent variant alleles (**Supplementary Figure 25 and 26**) and found that 22,133,782 bubbles represented 20,194,117 SNP alleles, 6,848,115 indel alleles, and 413,809 SV alleles (Methods, **Supplementary Figure 27**). Of these non-reference SV alleles, 17,720 were observed in bi-allelic contexts and 396,089 at multi-allelic loci with more than one non-reference allele, including extreme cases where all 88 haplotypes showed distinct alleles (**Supplementary Figure 27**). In order to analyze PanGenie's genotyping performance, we conducted a "leave-one-out" experiment in which we repeatedly removed one sample from the graph and re-genotyped it using the remaining haplotype paths in the graph and short-read data for the left out sample (Methods). In line with previous results (Ebert et al., 2021); (Ebler et al., 2022), we obtained high genotype concordances across all variant types and genomic contexts (**Supplementary Figure 28**). Furthermore, we used PanGenie to genotype HG002 and evaluated genotypes based on SVs at challenging medically-relevant loci (Wagner, Olson, Harris, McDaniel, et al., 2022), resulting in a precision of 0.74 and an adjusted recall of 0.81 (Methods).

Next, we genotyped the 3,202 samples from the 1KG (Byrska-Bishop et al., 2021) (Methods). We filtered the resulting SV genotypes using a machine-learning approach (Ebert et al., 2021; Ebler et al., 2022) that assessed different statistics, including Mendelian consistency and concordance to assembly-based calls. As a result, we produced a filtered, high-quality subset of SV genotypes containing 28,434 deletion alleles, 84,752 insertion alleles, and 26,439 other SV

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

alleles (**Supplementary Table 16**, Methods). Expectedly, many of the alleles not included in the filtered set stem from complex, multi-allelic loci and are enriched for rare alleles. As independent quality control measures for genotypes in the filtered set, we assessed Hardy-Weinberg Equilibrium (**Supplementary Figures 29, 30, and 31**) and compared allele frequencies observed across the genotypes of all 2,504 unrelated samples to the respective allele frequencies of the 44 assembly samples contained in the graph and observed correlations (pearson) of 0.96, 0.93 and 0.90, respectively (**Figure 6C**), indicating high-quality genotypes. To quantify our ability to detect additional SVs, we compared our filtered set of genotypes to the HGSVC PanGenie genotypes (v2.0 “lenient” set, (Ebert et al., 2021)) and Illumina-based 1KG SV genotypes (Byrska-Bishop et al., 2021). The HGSVC and HPRC call sets are based on running PanGenie for re-genotyping variant calls produced from haplotype-resolved reference assemblies of disjoint sets of 32 and 44 samples, respectively, while the 1KG call set contains short-read based variant calls produced for each of the 3,202 1000 Genomes samples. In order to compare the call sets despite these differences, we analyzed the number of detected SV alleles in each sample (homozygous or heterozygous) and stratified by genome annotations from GIAB (**Figure 6D**, Methods) as well as using our own more detailed annotations **(Supplementary Figure 32)**. Results show that both PanGenie-based call sets detect more SVs (HPRC: 18,483 SVs/sample, HGSVC: 12,997 SVs/sample) than the short-read-based 1KG call set (9,596 SVs/sample), with an especially pronounced advance for deletions < 300 bp and insertions (**Figure 6E**). The respective average numbers of SVs per haplotype are 12,439 for HPRC, 9,227 for HGSVC and 6,099 for the 1KG calls (**Supplementary Figure 33**); that is, a gain of 104.0% HPRC over 1KG and of 34.8% over HGSVC. This confirms that short-read based SV discovery relative to a linear reference genome misses a large portion of SVs (Chaisson et al., 2019; Ebert et al., 2021; Zhao et al., 2021). Expectedly, the number of SVs per sample within “easy” genomic regions is consistent across all three callsets, while especially in low mappability and tandem repeat regions, the use of our pangenome reference leads to pronounced gains (**Figure 6D**), including for common variants **(Figure 6E, Supplementary Figure 34)**.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

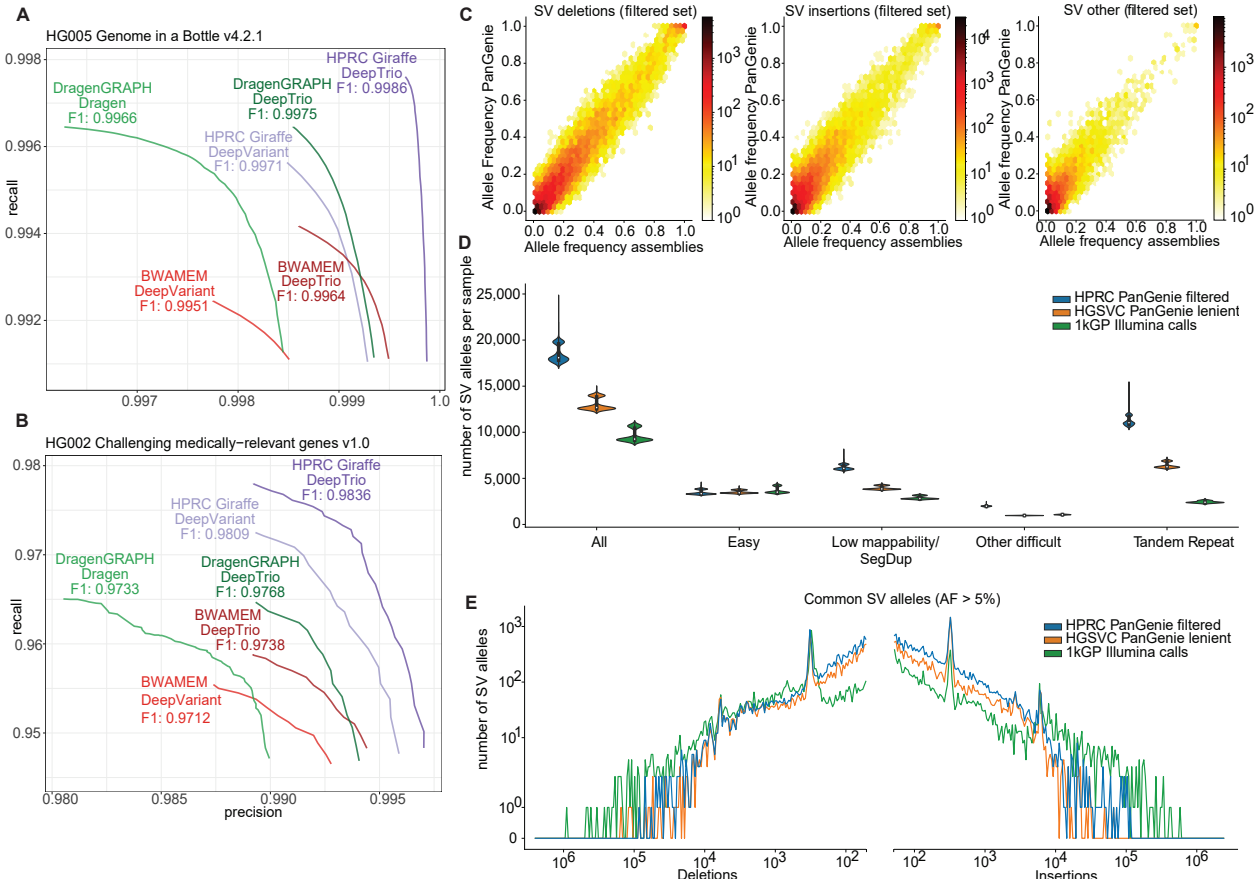


Figure 6 | Performance gains for pangenome-aided analysis of short-read whole-genome sequencing data. A-B) Precision-recall curves showing performance of different combinations of linear reference and various mappers and variant callers evaluated against (**A**) the GIAB v4.2.1 HG005 benchmark and (**B**) the challenging medically-relevant genes (CMRG) v1.0 benchmark. Giraffe uses the MC graph, BWA MEM uses GRCh38 and DragenGRAPH uses GRCh38 with additional alternative haplotype sequences. **C)** Comparison of allele frequencies observed from the PanGenie genotypes for all 2,504 unrelated 1KG samples and the allele frequencies observed across the 44 assembly samples in the MC graph. The PanGenie genotypes include all variants contained in the filtered set (28,433 deletions, 84,755 insertions, 32,431 other alleles). **D)** Shown are the number of SVs present (genotype 0/1 or 1/1) in each of the 3,202 1KG samples in the filtered HPRC genotypes (PanGenie), the HGSVC lenient set and the 1KG Illumina calls in GIAB regions. **E)** Shown is the length distribution of SV insertions and SV deletions contained in the filtered HPRC genotypes (PanGenie), the HGSVC lenient set and the 1KG Illumina calls. Only variants with an allele frequency > 5% across the 3,202 samples are considered.

Improved representation of tandem repeats

VNTRs are particularly variable between individuals and challenging to access with short-reads. The gains in the number of genotypable SVs in VNTRs (**Figure 6D, Supplementary Figure 34**) prompted us to investigate whether our pangenome reference would also improve read mapping in VNTR regions. To facilitate such an evaluation, we first established an orthology mapping between haplotypes in our pangenome reference using danbing-tk (Lu et al., 2021). The orthology can be established for 94,452 out of the 98,021 VNTR loci (96.4%)

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

discovered by TRF (Benson, 1999). We then simulated paired-end error-free short-reads from each genome at ~30x coverage. When mapping to GRCh38 with BWA-MEM, the rate of unmapped reads is 6.6-8.5 times greater compared to mapping to the MC graph with vg giraffe (**Figure 7A, Supplementary Table 17, Supplementary Figure 35**). The graph approach also outperformed the alternative in terms of true positives (TP), true negatives (TN), and false negatives (FN) (**Figure 7A**): The TN was on average 1.9% higher than the GRCh38 approach, and the TP was on average 0.087% higher. The graph approach also reduced FN by 2.1 fold. The slight increase in FP is possibly due to the boundary annotation of VNTRs on assemblies.

Given that the read depth over a locus is correlated with the copy number of a duplication, or the length of a tandem duplication, we evaluated how well length variants in VNTR regions can be estimated using either the MC graph or GRCh38. Using r^2 to measure how well the estimated lengths correlate with the actual lengths, the results showed that the graph approach performed better for 80% of the loci (48,085/60,386) and increased the median r^2 from 0.58 to 0.70 (**Supplementary figure 36**).

Improved RNA-seq mapping

To evaluate the benefit of our pangenome reference on transcriptomics, we simulated RNA-seq reads and evaluated the gains from mapping to a pangenomic reference compared to a standard reference genome (Methods). We included a simple pangenome model based on previous 1KG variant calls for comparison purposes (Byrska-Bishop et al., 2021; Sibbesen et al., 2021). We observed a notable gain in precision of RNA-seq mapping when to the MC graph augmented with splice junctions compared to the one derived from the 1KG pangenome graph. Both pangenome-based pipelines achieve significantly lower false mapping rates than a linear-reference pipeline using STAR (Dobin et al., 2013) (**Figure 7B**). With real sequencing data, mapping rates are more difficult to interpret in the absence of a ground truth (**Supplementary Figure 37**). Instead, we focus on the correlation in exon coverage to independent Iso-Seq data and find that the correlation is highest when mapping to a spliced pangenome graph derived from the MC graph (**Supplementary Figure 38**). The increase in correlation over the spliced pangenome graph derived from the 1000GP is modest but consistent across mapping quality (MAPQ) thresholds.

Improved ChIP-seq analysis

We used the pangenome to re-analyze H3K4me1 and H3K27ac ChIP-seq and ATAC-seq data from monocyte-derived macrophages obtained from 30 individuals (Groza et al., 2022). Overall, we observed a net increase in the number of peak calls, where, on average, 2 to 3% of peaks were found only when using the MC pangenome (**Figure 7C**). Moreover, the newly found peaks were replicated in more samples than expected by chance (**Supplementary Figure 39**). Additionally, we also used the pangenome to recover epigenomic features that were specific to alleles of SVs. For example, across all H3K4me1 samples, we assigned 1326 events to the SV allele, 1443 to the reference allele, and 2008 events to both alleles within heterozygous SVs (**Figure 7D**), with some replicated multiple times across samples (**Supplementary Figure 40**).

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

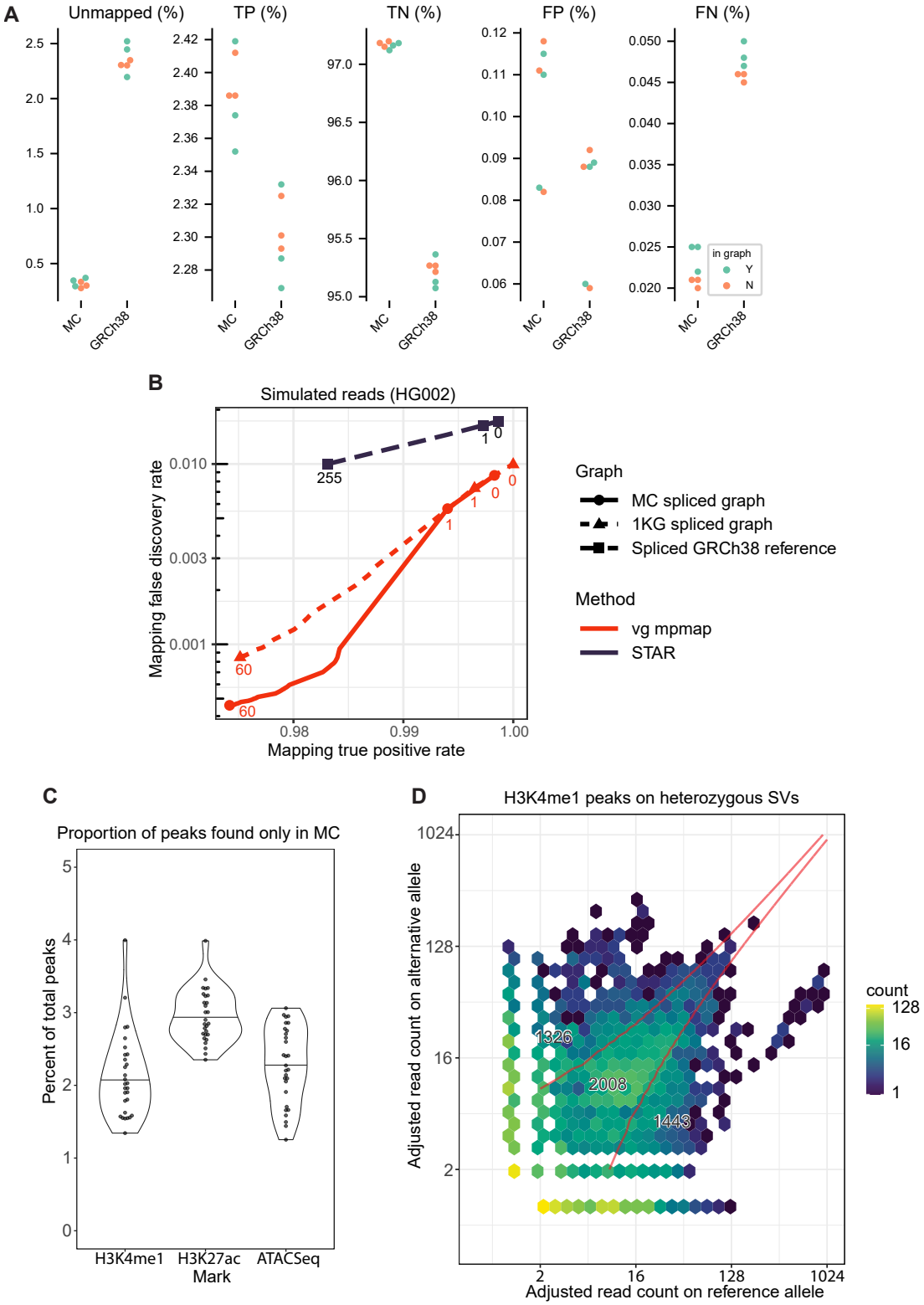


Figure 7 | Additional applications supported by the pangenome reference. A) Performance of read alignment in VNTR regions using the MC graph versus GRCh38. All statistics are expressed relative to the total number of reads simulated from each genome. **B)** Performance of RNA-seq read alignment. Mapping rate and false discovery rate are stratified by mapping

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

quality producing the curves shown. The MC graph is compared to a graph derived from the 1KG variant calls and to GRCh38. Each reference is augmented with splice junctions. vg mpmap was used to map to the graphs, and STAR was used to map to the linear reference. **C)** Proportion of all ChIP-seq peaks that are called only in the MC graph. Each data point represents samples that were assayed for H3K4me1, H3K27ac histone marks or chromatin accessibility using ATAC-seq. **D)** H3K4me1 peaks that overlap an SV for which the sample is heterozygous. The reads within the peak are partitioned between the SV or reference allele. The red boundary represents regions where a binomial test assigns a peak to the SV allele, both alleles, or the reference allele.

Discussion

We have publicly released 94 *de novo* haplotype assemblies from a diverse group of 47 individuals. This is the largest set of fully phased genome assemblies currently available, outperforming earlier efforts on many levels of assembly quality (Audano et al., 2019; Ebert et al., 2021; Shafin et al., 2020). For example, compared to Ebert et al., which was based primarily on more error-prone CLR instead of HiFi sequence data, the average median base-level accuracy is nearly an order of magnitude higher, the N50 measured contiguity of the phased assemblies is nearly double, and the assemblies are substantially more structurally accurate (see Porubsky et al. companion). These improvements are the result of recent major improvements in *de novo* assembly driven both by better sequencing technology and coordinated innovations in assembly algorithms (Cheng et al., 2021; Jarvis et al., 2022). To validate assembly structural accuracy we developed a new pipeline that maps low-error, long-reads to each diploid assembly to support the predicted haplotypes. This pipeline indicates more than 99% of each assembly, and greater than 90% of the assembled sequence representing highly repetitive arrays, is structurally correct, even though some challenges around difficult loci harboring copy-number polymorphisms and/or inversions remain (see Porubsky et al. companion). Highly accurate haplotype-resolved assemblies allow us to access previously inaccessible regions highlighting novel forms of genetic variation (as in Figure 5) and providing new insights into mutational processes such as interlocus gene conversion (see Vollger et al, companion).

Accompanying these assemblies are 94 sets of Ensembl gene annotations, representing the largest collection to date of *de novo* assembled human transcriptome annotations. Each transcriptome annotation is nearly complete, with fewer than 0.1% of GENCODE protein-coding genes and transcripts unannotated in each genome, and fewer than 0.2% of noncoding genes and transcripts missing. These putative transcriptome annotations allow us to analyze sequence-resolved copy-number variation; we assemble genic CNVs (mostly singletons) for 1529 different protein-coding genes, confirming earlier mapping-based analyses that predict the majority of rare genic CNVs occur outside of known SDs (Sudmant et al., 2010). These CNV genes account for 0.6-8.4 Mb of additional genic sequences per haplotype compared to GRCh38. These contain genes known to have CNV associated with human health including amylase (Falchi et al., 2014) (4-10 copies), beta-defensin (Mohajeri et al., 2016) (3-8 copies, *DEFB107A*), and NOTCH2NLC/B (Fiddes, Lodewijk, et al., 2018) (1 additional copy).

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The pangenomes presented are both a set of individual haploid genome assemblies and an alignment of these assemblies. The combination can be efficiently and elegantly described as a variation graph (Eizenga et al., 2020, 2021). A new set of exchange formats for pangenomics, including extensions of Graphical Fragment Format (GFA) that encode variation graphs, are emerging (Li et al., 2020). In a companion to this work, Siren et al. demonstrate that the pangenomes presented here can be losslessly stored using a compressed, binary representation of GFA in just ~3-6 gigabytes (Sirén & Paten, 2022) despite representing more than 282 billion bases of individual sequence, with strongly sublinear scaling as new genomes are added. Creating pangenome graphs is an active research topic so we developed multiple pipelines, with details of these methods further explored in companion papers (Garrison, Guarracino, et al., 2022; Hickey et al., 2022). We demonstrate concordance between these different construction approaches; the MC and PGGB pangenomes contain nearly the same number of small variants and SVs of various types. Further, these encoded pangenome variants show high levels of agreement with existing linear reference-based methods for variant discovery, particularly within the non-repetitive fraction of the genome. Where the pangenome drafts presented differ is principally in how they handle CNV sequences. The PGGB method will frequently merge copies of a CNV, while the MC graphs represent CNV copies as independent subgraphs. Both approaches have merits, and which approach to favor will take further experimentation and community input, and may vary by the specific application. The PGGB method retained all centromeric and satellite sequences, while the MC graph pruned much of this sequence. This made it practical with current methods to use the MC graphs for read alignment applications. However, pruning these sequences is not a satisfactory solution. Longer-term, more work is needed to determine how best to align and represent these large repeat arrays within pangenomes, particularly as T2T assembly becomes commonplace and these arrays are therefore finished. Furthermore, although the PGGB graph retained centromeric and satellite sequences, in principle enabling analysis of previously-inaccessible parts of the pangenome, our initial population-genetic analysis of these regions (Methods) leaves open questions about assembly accuracy and alignment especially in areas of the genome where mutation rates are thought to be an order of magnitude greater (Logsdon et al., 2021). This suggests that significant care must be taken when studying them, and new methods may need to be developed to fully understand and characterize this component of the human pangenome.

A near-term application of pangenome references will be to improve reference-based sequence mapping workflows. In these workflows, the pangenome can act as a drop-in replacement for existing references, with the read mappings projected from pangenome space back onto an existing linear reference for downstream processing. This is how the Giraffe-DeepVariant workflow functions: DeepVariant, the variant caller, never needs to consider the complexity of the pangenome, but the workflow benefits from a mapping step that accounts for sequences that are missing from the linear reference. Making the switch to using pangenome mapping is not significantly more expensive computationally (Sirén et al., 2021), and resulted in an average 34% reduction in errors vs. using the standard reference methods (**Supplementary Figure 41**). These benefits were also greatest at complex loci: for example, we found the largest absolute increases in accuracy using the GIAB challenging medically relevant genes benchmark

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

(Wagner, Olson, Harris, McDaniel, et al., 2022). Pangenomes do not just improve variant calling: we build on recent work to further show that mapping to the pantranscriptome graph similarly improves transcript mapping accuracy (Sibbesen et al., 2021), and pangenome mapping can help improve the detection of ChIP-seq peaks (Groza et al., 2020).

SVs have been mostly excluded from short-read studies because methods to genotype them using the linear reference have limited accuracy and sensitivity. Previous short-read, linear reference studies have discovered 7.5-9.5k SVs per sample (Byrska-Bishop et al., 2021; Collins et al., 2020) while long read-sequencing efforts have routinely discovered ~25k. Ebert et al. showed that using PanGenie, a pangenomic approach, with 32 samples a subset of these variants could be genotyped in short read genomes (~13k genotyped on average, ranging from 12.0k to 15.0k per sample). Using the same PanGenie method, the HPRC pangenome increases this to ~18.5k (ranging from 16.9k to 24.9k) per sample using the same method, allowing the genotyping of the substantial majority of SVs discovered using long-reads per sample. The draft pangenome therefore delivers much better SV calling than earlier approaches, extracting latent information from short-read samples that are already available, and so in the future enabling the inclusion of tens of thousands of additional SV alleles into genome-wide association studies (GWAS). Relative to Ebert et al., it is likely this improvement is a combination of advances in sequencing (HiFi) and assembly, increased numbers of individuals in the pangenome and the full sequence-level representation of SV alleles in the pangenome graph, avoiding merging of similar but distinct alleles. Looking beyond short-reads, in the future, the combination of the pangenome and low-cost long-read sequencing should prove a potent combination for comprehensive SV genotyping.

The openly accessible, diverse assemblies and pangenome graphs we present today form a draft of a pangenome reference. There are many remaining challenges in growing and refining this reference. For example, assembly reliability analysis revealed roughly an order of magnitude more erroneously assembled sequences in the HPRC assemblies than in the T2T-CHM13 complete assembly. Furthermore, despite being predicted to have less than one base error per two hundred thousand assembled bases, base-level sequencing errors are still an issue. For example, in line with this error rate, we identified more than a dozen apparent frameshifts and nonsense mutations per genome annotation that are likely the result of sequencing errors. The cohort we present is also relatively small notwithstanding the significant effort to generate the underlying long-read sequencing resource. Our near-term goal is to expand the pangenome to a diverse cohort of 350 individuals (which should capture most common variants), to push toward T2T genomes for this cohort (to properly represent the entire genome in almost all individuals), and to refine the pangenome alignment methods (so that telomere-to-telomere alignment is possible capturing more complex regions of the genome). This will give us a dramatically more comprehensive representation of all types of human variation.

We acknowledge that references generated from 1KG samples alone are insufficient to capture the extent of sequence diversity in the human population. To ensure that we are able to maximize our surveys of sample diversity, while abiding by principles of community engagement

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

and avoiding extractive practices (Eizenga et al., 2020; Wang et al., 2022), we will broaden our efforts to recruit new participants to improve the representation of human genetic diversity. A richer human reference map promises to improve our understanding of genomics and our ability to predict, diagnose and treat disease. A more diverse human reference map should also help to ensure that the eventual applications of genomic research and precision medicine are effective for all populations. We acknowledge that the value of this project will partly be in the future establishment of new standards for how we capture variant diversity, the opportunity to disseminate science into diverse communities, and continued efforts to engage with diverse voices in this ambitious goal to build a common global reference resource. In parallel with our efforts to reach a more comprehensive collection of diverse and highly accurate human reference genomes, we expect further optimization and rapid improvement of the pangenome reference, enabling an increasingly broad set of applications and use cases for both the research and clinical communities.

Methods

Sample Selection

We identified 1KG parent-child trios in which the child cell line banked within NHGRI Sample Repository for Human Genetic Research at the Coriell Institute for Medical Research was listed as having zero expansions and 2 or fewer passages, and rank-ordered representative individuals as follows. Loci with minor AF (MAF) less than 0.05 were removed. MAF was measured in the full cohort (i.e. 2504 individuals, 26 subpopulations) regardless of each individual's subpopulation labeling. For each chromosome, Principal Component Analysis (PCA) was performed for dimension reduction. This resulted in a matrix with 2200 features, which was then centered and scaled with smartPCA normalization. The matrix was further reduced to 100 features through another round of PCA.

We defined the representative individuals of a subpopulation as those who are similar to the other members in the group (which, in this scenario, is the subpopulation they belong to), as well as different from individuals outside the group. Group is defined by previous 1KG population labels (e.g "Gambian in Western Division"). We do this as follows. For each sample, we first calculate the intra-group distance d_{intra}, which is the average of L2-norms between the sample and samples of the same subpopulation. Inter-group distance, d_{inter}, is similarly defined as the average of L2-norms between the sample and samples from all other subpopulations. The L2-norms are derived in the PCA's feature space. The score of this sample is then defined as $10 \times d_{intra} + d_{inter}/(n - 1)$, where n is the number of subpopulations. For each subpopulation, if fewer than three trios are available, all are selected. Otherwise, trios are sorted by ranking children with $max(paternal_{rank}, maternal_{rank})$, where $paternal_{rank}$, and $maternal_{rank}$ are the respective ranks of each parent's score, selecting the three trios with maximum value. We ranked by parent scores because during the YR1 effort the child samples did not have sequencing data and therefore had to be represented by the parents.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

At this point, ideally, we would like to select the same number of candidates from each subpopulation, and have an equal number of candidates from both genders. To correct for imbalances, we applied the following for each subpopulation's candidate set: a) when gender is unbalanced (i.e. off by more than one sample), we tried to swap in the next-best candidate of the less represented gender; do nothing if this is not possible. b) if a subpopulation has less individuals than the desired sample selection size (i.e. all candidates are selected), their unused slots will be distributed to other unsaturated subpopulations. The latter choice is arbitrary but should have little impact on the overall results.

Sequencing

Cell line expansion and banking for sequencing

Lymphoblastoid cell lines used for sequencing from the 1KG collection (**Supplementary Table 1**) were obtained from the NHGRI Sample Repository for Human Genetic Research at the Coriell Institute for Medical Research. HG002 (GM24385) and HG005 (GM24631) lymphoblastoid cell lines were obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research. All expansions for sequencing were derived from the original expansion culture lot to ensure the lowest possible number of passages and to reduce overall culturing time. Cells used for HiFi, Nanopore, Omni-C, Strand-seq, 10x Genomics, and Bionano production as well as g-banded karyotyping and Illumina Omni2.5 microarray were expanded to a total culture size of $4x10^8$ cells, resulting in a total of five passages post-cell line establishment. Cells were split into production-specific sized vials: HiFi ($2x10^7$ cells), Nanopore ($5x10^7$ cells), Omni-C ($5x10^6$ cells), Strand-seq ($1x10^7$ cells), 10x Genomics ($4x10^6$ cells), and Bionano ($4x10^6$ cells). Cells for Strand-seq were stored in 65% RPMI-1640, 30% FBS, and 5% DMSO and frozen as viable cultures. All other cells were washed in PBS and flash frozen as dry cell pellets. Cells used for ONT-UL production were separately expanded from the original expansion culture lot to a bank of five vials of $5x10^6$ cells. A single vial was subsequently expanded to a total culture size of $4x10^8$ cells, resulting in a total of eight passages. Cells were also reserved for g-banded karyotyping and Illumina Omni2.5 microarray.

Karyotyping and microarray

G-banded karyotype analysis was performed on $5x10^6$ cells harvested at passage 5 (for HiFi, Nanopore, and Omni-C) and passage 8 (for ONT-UL). For all cell lines, twenty metaphase cells were counted, and a minimum of five metaphase cells were analyzed and karyotyped. Chromosome analysis was performed at a resolution of 400 bands or greater. A pass/fail criteria was used before cell lines proceeded to sequencing. Cell lines with normal karyotypes (46,XX or 46,XY) or lines with benign polymorphisms that are frequently seen in apparently healthy individuals were classified as passes. Cell lines were classified as failures if two or more cells harbored the same chromosomal abnormality. DNA used for microarray was isolated from frozen cell pellets ($3x10^6$-$7x10^6$ cells) using the Maxwell RSC Cultured Cells DNA Kit on a Maxwell RSC 48 instrument (Promega). DNA was genotyped at the Children's Hospital of

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Philadelphia's Center for Applied Genomics using the Infinium Omni2.5-8 v1.3 BeadChip (Illumina) on an iScan System instrument (Illumina).

HiFi Sequencing

Pacific Bioscience HiFi sequencing was distributed between two centers, Washington University in St Louis (WashU) and the University of Washington (UW). We describe the protocols used at each center separately.

HiFi Production (Washington University in St Louis)

High-molecular-weight DNA was isolated from frozen cell pellets using Qiagen MagAttract HMW DNA kit and sheared using Diagenode Megaruptor I to 20 kb mode size. At all steps, DNA quantity was checked on the Qubit Fluorometer I with the dsDNA HS Assay Kit (Thermo Fisher) and sizes were examined on FEMTO Pulse (Agilent Technologies) using the Genomic DNA 165 kb Kit. SMRTbell libraries were prepared for sequencing according to the protocol 'Procedure & Checklist – Preparing HiFi SMRTbell Libraries using the SMRTbell Express Template Prep Kit 2.0'. After SMRTbell generation, material was size-selected on a SageELF system (Sage Science) using the "0.75% 1-18kb" program (target 3450 bp in well 12) and some combination of fraction3 (average size 15-21 kb), fraction 2 (average size 16-27 kb), and fraction 1 (average size 20-31 kb) were selected for sequencing, depending on empirical size measurements and available mass. The selected library fractions were bound with Sequencing Primer v2 and Sequel II Polymerase v2.0 and sequenced on Sequel II instruments (PacBio) on SMRT Cells 8M using Sequencing Plate v2.0, diffusion loading, two hour pre-extension, and 30 hour movie times. Samples were sequenced to a minimum HiFi data amount of 108.5 Gbp (35X estimated genome coverage) on four SMRT Cells.

HiFi Production (University of Washington)

High-molecular-weight DNA was isolated from frozen cell pellets using a modified Gentra Puregene method and sheared using gTUBE (Covaris, Inc.) to 20 kb mode size. At all steps, DNA quantity was checked by fluorometry on the DS-11 FX instrument (DeNovix) with the Qubit dsDNA HS Assay Kit (Thermo Fisher) and sizes were examined on FEMTO Pulse (Agilent Technologies) using the Genomic DNA 165 kb Kit. SMRTbell libraries were prepared for sequencing according to the protocol 'Procedure & Checklist – Preparing HiFi SMRTbell Libraries using the SMRTbell Express Template Prep Kit 2.0'. After SMRTbell generation, material was size-selected on a SageELF system (Sage Science) using the "0.75% 1-18kb" program (target 3400 bp in well 12) and fraction 2 (average size 17-20 kb) or fraction 1 (average size 18-20 kb) were selected for sequencing, depending on empirical size measurements and available mass. For some samples, the SageELF program "0.75% agarose, 10 kb-40 kb" (target 10000 bp in well 10) was used and fractions 6 and 7 were pooled together for sequencing (average size 17-21 kb). The selected library fractions were bound with Sequencing Primer v2 and Sequel II Polymerase v2.0 and sequenced on Sequel II instruments (PacBio) on SMRT Cells 8M using Sequencing Plate v2.0, diffusion loading, three- to four-hour pre-extension, and 30 hour movie times. Samples were sequenced to a minimum HiFi data amount of 96 Gbp (30X estimated genome coverage) on at least four SMRT Cells.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

HiFi Production methods comparisons

Although subtle differences in HiFi data production methods exist between WashU and UW, the resulting data was remarkably similar with overlapping assembly statistics from most samples. These initial genomes were sequenced at a time when methods were being refined and optimized for HiFi sequencing, as it was a relatively new process. The primary differences in protocols are part of the nucleic acid isolation, fragmentation, and size selection, with the downstream sequencing specific applications being much more consistent. Both teams were closely engaged with each other as well as with our company associates to provide optimal end products.

Nanopore Ultra-long sequencing protocol

For the HPRC+ samples we used the nanopore unsheared long-read sequencing protocol (Shafin et al., 2020). This generated ~60x coverage of unsheared sequencing from 3 PromethION flow cells, N50 ~44kb. For the HPRC samples we used the following protocol.

DNA Extraction

A 50 million cell pellet was resuspended in 200 µL of PBS and the resuspended cells aliquoted (40 µl) into 5x 1.5 ml DNA Lo-bind Eppendorf tubes. The following procedure for DNA extraction was completed for each of the 5 aliquots. Each tube contained sufficient DNA for 3 libraries loaded on 1 flow cell. In order, the following were added to each tube with pipette mixing (10X up and down) using a P200 wide-bore pipette: 40 µL of Proteinase K, 40 µL of Buffer CS; and 40 µL of CLE3. The samples were then incubated at RT (18–25 °C) for 30 min. Next, 40 µL of RNase A were added to each tube with pipette mixing (10X) with a P200 wide-bore and then samples were incubated at RT for 3 min. Two hundred microliters of BL3 were mixed with 200 µL PBS in a 1.5 mL Eppendorf tube. Four hundred microliters of this BL3/PBS mixture were then added to each sample and the samples pipet mixed 10X with a P1000 wide-bore pipette set to 600 µL.

Samples were incubated for 10 minutes at RT and then pipette mixed 5X, then incubated at RT for 10 mins and pipette mixed 5X and then further incubated for 10 minutes at RT. A white precipitate may form after addition of BL3. This is completely normal. A Nanobind disk is added to the cell lysate first then 600 µL of isopropanol and mixing is by inversion of the tube 5X. Tubes were further mixed on the tube rotator (9 rpm at RT for 10 min). The tubes were then placed on the magnetic tube rack and the nanobind disk positioned closer to the top of the tube to avoid inadvertent removal of the DNA bound to the nanobind disk. The supernatant was discarded with a pipette and 700 µL of Buffer CW1, was added to each tube. The tube in the magnetic rack is then inverted 4X for mixing. A second and third wash with 500 µL of Buffer CW2, (inversion mix 4X for each wash) was performed. After the 2nd CW2 wash, liquid was removed from the tube cap and the tubes spun on a mini-centrifuge for 2 s, and replaced on the magnetic rack. Residual liquid was removed from the bottom of the tube taking care to not remove DNA associated with the nanobind disk. Elution from the nanobind disk was accomplished by adding 160 µL Circulomics EB + 0.02% Triton-X100 (make by mixing 316.8 µL EB + 3.2 µL 2% Triton-X100) and incubation at RT for at least 1 hour. Tubes were gently tapped

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

half way through elution. DNA was collected by transferring eluate with a P200 wide-bore pipette to a new 1.5 mL microcentrifuge tube. Some liquid and DNA remain on the Nanobind disk after pipetting. We spun the tube containing the Nanobind disk on a centrifuge at 10,000 x g for 5 s and also transferred any additional liquid that came off the disk to the eluate tube. This process was repeated if necessary until all DNA was removed. The samples were pipette mixed 5X (approx. 10 seconds to aspirate and 10 seconds to dispense for each cycle) with a wide bore P200 pipette in order to homogenize the sample. Samples were further allowed to rest at RT overnight to allow DNA to solubilize (disperse).

Library Preparation, DNA Tagmentation / FRA

Circulomics EB+ (EB buffer with 0.02% Triton-X100) was prepared and 140.82 µL EB + aliquoted into a 1.5 ml Eppendorf DNA Lo-Bind tube. UHMW DNA (300 µL) from above was aliquoted into the same tube with a wide bore P200 pipette. The mixture was slowly pipetted up and down 3X with a wide bore P200 pipette set to 150 µL. In a separate 1.5 ml Eppendorf DNA Lo-Bind tube, in order, 144 µL of FRA Dilution Buffer, was added, 9.18 µL of 1 M MgCl2 was added, and 6 µL FRA was added. The tube was tapped to mix and spun down using a microcentrifuge. The EB/Triton/DNA mixture was added to the FRA Dilution Buffer/MgCl2/FRA mixture with a wide bore P200 pipette. This mixture was then pipette mixed 15-20X with a wide bore P1000 pipette set to 600 µL. The mixture appeared homogeneous when pipette mixing finished. The tube was then incubated for 15 min at RT. The mixture was then pipette mixed 5X with a wide bore P1000 pipette set to 600 µL and incubated at RT for an additional 15 mins. The mixture was then incubated at 30 °C for 1 min, followed by 80 °C for 1 min, and then held at 4 °C.

Library Preparation, FRA Reaction Cleanup

Cleanup employed a Nanobind disk. A 5 mm Nanobind disk was added to the reaction mixture above followed by 300 µL of Circulomics Buffer NAF10. The tube was gently tapped 10–20X to mix. The mixture was placed on a platform rocker at 20 rpm for 2 min at RT. A DNA “cloud” was visible on the Nanobind disk. The tube was spun for 1 – 2 sec using a benchtop microcentrifuge and placed on a magnetic rack. The binding solution was removed and discarded. The Nanobind disk was washed by adding 350 µL ONT Long Fragment Buffer (LFB) and gentle tapping 5X to mix. The tube was spun for 1–2 sec using a microcentrifuge and placed on a magnetic rack. The ONT Long Fragment Buffer (LFB) was removed and discarded. Care was taken to not pipette DNA attached to the Nanobind disk. This LFB wash was repeated. The tube was then briefly spun (microcentrifuge) to move the Nanobind disk to the bottom of the tube. DNA was eluted from the Nanobind disk by addition of 125 µL of ONT Elution Buffer (EB) to the tube. The tube was incubated for 30 min at RT, then gently tapped 5X (mixing) and incubated for an additional 30 min at RT. Fluid was slowly aspirated 4X over the Nanobind disk before removing the eluate from the tube. The eluate was transferred to a new 1.5 ml Eppendorf DNA Lo-Bind tube using a wide bore P200 pipette. The eluate was then pipette mixed 2X with a wide bore P200 pipette.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Library Preparation, Adapter Attachment / RAP

The RAP adaptor was next added to the DNA preparation. To 120 ul of eluate (from above), 3 µL of ONT Rapid Adapter (RAP) was added. The mixture was pipette mixed 8X with a wide bore P200 pipette. The mixture was then incubated for 15 mins. at RT and then again pipette mixed 8X with a wide bore P200 pipette.

Library Preparation, RAP Reaction Cleanup with Nanobind

Final library cleanup removes unligated adaptor. Add 120 µL Circulomics Buffer EB to the 123 µL RAP reaction mixture from above. Slowly pipette mix 3X with a wide bore P1000 pipette set to 240 µL. Each aspiration should take ~10 sec and each dispense should take ~10 sec. A 5 mm Nanobind disk was added to the reaction mixture followed by 120 µL Circulomics Buffer NAF10. Mixing was accomplished by gentle tapping. The tube was incubated for 5 min at RT without agitation/rotation. Gently tap 5X 2 – 3 times during the 5 min incubation. The tube was spun for 1 – 2 sec using a microcentrifuge and placed on a magnetic rack. The binding solution was discarded. Next 350 µL of ONT Long Fragment Buffer (LFB) was added to the tube and mixed by gentle tapping 5X. The tube was then spun for 1 – 2 sec using a microcentrifuge and placed on a magnetic rack. The ONT Long Fragment Buffer (LFB) was removed and discarded. Next the Nanobind disk was washed by adding 350 µL ONT Long Fragment Buffer (LFB). The tube was gently tapped 5X to move LFB over the surface of the disk. The tube was then incubated at RT for 5 min. The tube was then spun for 1 – 2 sec using a microcentrifuge and placed on a magnetic rack. The ONT Long Fragment Buffer (LFB) was removed and discarded. The tube was briefly spun using a microcentrifuge to move the Nanobind disk to the bottom of the tube. To elute DNA from the Nanobind disk, 126 µL ONT Elution Buffer (EB) buffer was added to the tube. The tube was incubated for 30 min at RT, then gently tapped 5 – 10X and incubated for an additional 1-2 hours at RT. The eluate was then transferred to a new 1.5 ml Eppendorf DNA Lo-Bind tube using a wide bore P200 pipette using the same technique described above for passing the eluate over the Nanobind disk before removing the eluate from the tube. The mixture was then pipette mixed 2 – 3X with a wide bore P200 pipette. The library was stored overnight at 4 ºC prior to sequencing to permit maximal dissolution of DNA.

Flow Cell Loading and Sequencing

ONT Sequencing Buffer (SQB) (68 µL) was added to 82 µL of the eluate from above. The mixture was pipette mixed 4X with a wide bore P200 pipette set to 150 µL. Each aspiration of 150 µL should take ~10 – 20 sec, and each dispense of 150 µL should take ~10 – 20 sec. Samples were then incubated at RT for 10 min. Next the samples were again pipette mixed 8X with a wide bore P200 pipette set to 150 µL as above. Before loading the library, the flow cell was primed with flush buffer/flush tether mixture per Oxford Nanopore Technologies directions. The library was then added to the flow cell. The mixture was viscous, but loaded smoothly in about 1 min. Some samples took 2 mins max to load. The sequencing run had re-mux time set for every 6 hours. Basecalling was performed with Guppy version 4.0.11, using default parameters and the high-accuracy PromethION model (dna_r9.4.1_450bps_hac_prom.cfg).

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Dovetail Omni-C

We prepared Omni-C libraries from each cell line using the Dovetail Omni-C Kit (Dovetail Genomics, CA) with modifications as follows. First, we aliquoted 1 million cells for fixation with Formaldehyde and DSG. We digested chromatin with DNAse I until DNA fragments of a desired length were obtained. Per the protocol, we performed end repair on the chromatin, followed by the ligation of a biotinylated bridge oligo, followed by ligation of free chromatin ends. We reversed cross links and purified proximity ligated DNA. We converted the DNA into an Illumina sequencing library using the NEB Ultra II library preparation kit (NEB, Ipswich, MA) with a Y-adaptor. We enriched for ligation products using streptavidin bead capture on the final library. Each capture reaction was then split into two replicates prior to the final PCR enrichment in order to preserve complexity. All libraries were uniquely dual indexed and sequenced on an Illumina Novaseq Platform with read lengths of 2x150bp.

Phased Assembly Pipeline

We describe the main automated and manual steps taken before, during, and after assembly. A combined set of Workflow Description Language (WDL) formatted assembly workflows is available from Dockstore that captures each of the steps for filtering adapter-contained reads and running hifiasm, (https://dockstore.org/organizations/HumanPangenome/collections/Hifiasm). All assemblies were generated using this workflow, running on AnVIL (Schatz et al., 2022). Cleaning assemblies and fixing some structural issues were performed through a combination of automated workflows and manual curation described below.

Filtering adapter-contained reads and running Hifiasm

Before producing the assemblies we detected and removed the reads containing PacBio adapters, using a bash script from the HiFiAdapterFilt repository (Sim, 2021) (`commit:64d1c7b9f6511ed8934ed2faf09f301f459db43b`). This script first creates a database of the PacBio adapter sequences, as illustrated below:

```
>gnl|uv|NGB00972.1:1-45 Pacific Biosciences Blunt Adapter
ATCTCTCTCTTTTCCTCCTCCTCCGTTGTTGTTGTTGAGAGAGAT
>gnl|uv|NGB00973.1:1-35 Pacific Biosciences C2 Primer
AAAAAAAAAAAAAAAAAATTAACGGAGGAGGAGGA
```

It then runs `blastn` with tuned parameters to detect adapter-containing reads:

```
blastn -db ${DATABASE} -query ${HIFI_FASTA} -task blastn -reward 1 -penalty -5 -gapopen 3 -gapextend 3 -dust no -soft_masking true -evalue 700 -searchsp 1750000000000 -outfmt
```

For 43 samples (out of 47) we removed less than 0.15% of the reads; it is worth noting that all of the 29 HPRC samples are among these 43 samples, indicating the low level of adapter contamination in the HiFi data produced by the HPRC. HG005, which is an HPRC+ sample, had the highest contamination percentage, at ~1%. **(Supplementary Figure 42)**

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The removed reads were then aligned to the T2T-CHM13v2.0 reference to ensure that there is no chromosomal or locus-specific bias in the filtering process. **(Supplementary Figure 43)** shows a snapshot of the IGV browser (Robinson et al., 2017) illustrating the coverage of the adapter-containing reads along the genome. It shows that the locations of the reads are almost evenly distributed along the genome and, excluding centromeres, we barely find any region covered with more than 2 adapter-containing reads, even in HG005 which had the highest contamination percentage.

The trio-binning mode of Hifiasm needs haplotype-specific kmers for trio phasing the assembly graph. To generate these kmers we used parental Illumina short-reads for the 47 HPRC/HPRC+ samples, which are publicly available from the 1KG dataset (Byrska-Bishop et al., 2021). For each parental short-read sample we used `yak count` (v0.1, (Li, 2020)) to generate the kmer hash tables, running it once for each of the paternal and maternal read sets.

```
yak count -k31 -b37 -o pat.yak paternal.fq.gz
yak count -k31 -b37 -o mat.yak maternal.fq.gz
```

The adapter filtered HiFi reads along with the parental kmer tables were then given to Hifiasm-v0.14 to produce haplotype-resolved assembly graphs. Only the sample HG002 was re-assembled with Hifiasm-v0.14.1, which is explained in more detail in the next subsection.

```
hifiasm -o ${SAMPLE_NAME} -t 48 -1 pat.yak -2 mat.yak hifi.fq.gz
```

Hifiasm produces one graph per haplotype in GFA format. Each haplotype-specific GFA file is then converted to FASTA format using Gfatools (Li, 2021a). The assemblies produced by Hifiasm-v0.14 are released under v2 after doing the three cleaning steps described at the end of this section.

Fixing issues manually

We used `paftools.js asmgene`, from the minimap2 repository (https://github.com/lh3/minimap2/tree/master/misc) (Li, 2021b), to count the number of apparent gene duplications for each of the assemblies produced by Hifiasm-v0.14. This assessment acted as a proxy for detecting high-level duplication errors. We used the Ensembl v99 cDNA sequences (Cunningham et al., 2022) as the input gene set for running asmgene.

```
# Aligning genes to GRCh38 and each Hifiasm haploid assembly
minimap2 -cx splice:hq hs38.fa cdna.fa > hs38.paf
minimap2 -cx splice:hq ${pat/mat}.fa cdna.fa > ${pat/mat}.paf

# Detecting gene duplications
paftools.js asmgene -a hs38.paf ${pat/mat}.paf
```

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Three samples were detected as outliers in terms of the number of gene duplications. To identify the cause of this issue we aligned back the HiFi reads to those assemblies and checked the depth of coverages and mapping qualities. It showed that the samples HG01358, HG01123, and HG002 contained false duplications of length ~55Mb (in `h1tg000058l` contig), ~14Mb (`h1tg000013l`), and ~70Mb (`h2tg000045l`) respectively. In the assembly graphs of HG01358 and HG01123 the duplicated HiFi reads appearing multiple times were used as anchors to manually determine the exact boundaries of the duplicated regions in the contigs. These two contigs were then fixed manually by breaking the contigs at the duplication start and end points and discarding the duplicated sequence from the assembly. In detail, for HG01123 for `h1tg000013l` we discarded the interval `[94439457, contig end]`. For HG01358 for `h1tg000058l` we kept the interval `[0, 95732608)`, renaming the contig to `h1tg000058l_1`, we discarded the interval `[95732608, 150395342)` and kept the interval `[150395342, contig end]`, renaming it to `h1tg000058l_2`. To address the false duplication in HG002 we re-assembled it using a newer version of Hifiasm; v0.14.1, which was reported not to have this problem.

We also evaluated the phasing accuracy of the assemblies by using `yak trioeval` (see below). We detected a single large misjoin in a maternal contig of the HG02080 assembly. It contained a ~22Mb long paternal block in the middle of the contig and as a result two switch errors at the edges of this block. This block was manually discarded from the assembly and the contig was broken into two smaller ones. In detail, in HG02080 for the `h2tg000053l` contig we kept the interval `[0, 41506503)`, renaming it to `h2tg000053l_1`, we discarded the interval `[41506503, 63683095)`, and kept the interval `[63683095, contig end]`, renaming it to `h2tg000053l_2`.

We finally searched for interchromosomal misjoins using the minigraph pangenome (see below for construction details). An “interchromosomal misjoin” was defined by a chimeric minigraph alignment (see below) consisting of ≥1Mb sub-alignments on different chromosomes.

Cleaning steps

To clean the raw assemblies we performed three additional steps. In summary, these steps consisted of masking the remaining HiFi adapters, dropping the contigs that were contaminated in their entirety, and removing any redundant mitochondrial contigs.

In the first cleaning step, the sequence of PacBio SMRTbell adapter was aligned to each assembly using `minimap2` with the parameters ‘`-cxsr -f5000 -N2000 --secondary=yes`’. We extracted only the hits with less than or equal to 2 mismatches and which were longer than 42nt. In addition, eukaryotic adapters in each assembly were identified by `VecScreen` (*VecScreen: Screen for Vector Contamination*, n.d.). The combined `minimap2`

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

and `VecScreen` adaptor hits (when present) were hard-masked in the assemblies using a WDL of the bedtools maskfasta command (https://dockstore.org/workflows/github.com/human-pangenomics/hpp_production_workflows/MaskAssembly:master?tab=info).

```
bedtools maskfasta \
      -fi ${inputFastaFN} \
      -bed ~{adapterBed} \
      -fo ~{outputFasta}
```

In the second cleaning step, we used `VecScreen` to detect mitochondrial contigs and the contigs consisting non-human sequences from other organisms like bacteria, viruses, and fungi. These contigs were then dropped from the assemblies using a WDLized version of samtools faidx. It is worth noting that the contigs with nuclear mitochondrial DNAs within them were not dropped.

```
samtools faidx \
      $inputFastaFN \
      `cat contigsToKeep.txt` | gzip \
      > ~{outputFasta}
```

In the last cleaning step, we selected one contig as the best mitochondrial contig per diploid assembly. To do this selection we aligned the sequence of the mitochondrial DNA (with the RefSeq ID of NC_012920.1) to each diploid assembly using `minimap2` with the parameters '`-cx asm5 --cs`'. Then we selected one contig with the highest mapping score and the lowest number of mismatches as the best mitochondrial contig (we selected one randomly if multiple best contigs exist). This contig was then rotated and flipped (if neccesary) to match the start and orientation of NC_012920.1.fa and then added to the maternal assembly of the corresponding sample. Only the HG01071 sample did not produce any identifiable mitochondrial contig.

Masked, cleaned, mito assemblies were then accessioned to Genbank where they underwent another round of adapter masking and removal of (mostly EBV) contamination. The final assemblies were downloaded from Genbank and the contig IDs were pre-pended with the sample name and haplotype integer (where 1=paternal, and 2=maternal). For example, a contig assigned the name JAGYVH010000025 in sample HG02257's maternal assembly was renamed to be HG02257#2#JAGYVH010000025. The renamed assemblies were then released to our S3 and GCP buckets. In the process of download from Genbank, three of the assemblies (HG00733 paternal, HG02630 paternal, NA21309 maternal) had their downloads prematurely stopped resulting in missing sequence. Notably, NA21309 is missing its mitochondrial contig. Details can be found on the HPRC's Year 1 Assembly GitHub repository (https://github.com/human-pangenomics/HPP_Year1_Assemblies). The assemblies held in

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

INSDCs are not truncated, but the truncated copies were retained in S3 and GCP as they were used in construction of the pangenomes.

After submission to Genbank, the assemblies were aligned against CHM13 using Winnowmap and multiple contigs were found to be unmapped. These contigs were BLASTed and found to be almost exclusively EBV sequence. Genbank confirmed [personal communication] that these unmapped contigs should have been dropped as contamination. A list of the contigs that should have been dropped can be found on the Y1 Assembly GitHub repository (https://github.com/human-pangenomics/HPP_Year1_Assemblies/blob/main/genbank_changes/y1_genbank_remaining_potential_contamination.txt).

Assembly Assessment Pipeline

Several steps in assembly assessment were managed through a `StandardQC` workflow written using Workflow Description Language (WDL), run on AnVIL, and available in Dockstore (https://dockstore.org/workflows/github.com/human-pangenomics/hpp_production_workflows/StandardQC). Individual tools within the workflow were run in Docker containers with specific tool versions installed for consistency and reproducibility. Details are available within the Dockstore deposited workflow. The `StandardQC` workflow takes short-read data for parental and child samples, the two assembly haplotypes, and it produces an analysis over various quality metrics produced by the tools described below. For each task, the workflow produces a small human-readable summary file, which is also easy to parse for summarizing steps, as well as the full output from the tool for manual inspection. Specific tool invocations can be determined from the deposited workflow and are also described in the following sections.

Measuring potential interchromosomal joins

Contigs are aligned to CHM13 v2.0 with minigraph v0.18 and processed with the following command line:

```
minigraph -cxasm chm13v2.0.fa contigs.fa | paftools.js misjoin -
```

The “misjoin” command reports an interchromosomal join if a contig has two ≥1Mb alignments to two different chromosomes, respectively.

Assembly Contiguity Assessment

Assembly contiguity was assessed for each haplotype using QUAST (Mikheenko et al., 2018). These statistics include total sequence assembled, total assembled contigs, and contig NG50 (assuming a genome size of 3.1 Gb). All reference-based analyses were skipped.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

QUAST was invoked with the following command:

```
python /opt/quast/quast-5.0.2/quast-lg.py -t 16 -o <sample>.quast  --
large --est-ref-size 3100000000 --no-icarus
```

Assembly Quality Value Assessment

Assembly Quality Value (QV) was determined using two separate k-mer based tools. The first is yak (Cheng et al., 2021). Yak's QV estimation happens separately on each haplotype. Kmer databases for yak are generated with the following command:

```
yak count -t16 -b37 -o <sample>.yak <(cat <read_files>) <(cat
<read_files>)
```

QV estimation with yak is generated with the following command:

```
yak qv -t 32 -p -K 3.2g -l 100k <sample>.yak
<sample_assembly_haplotype> > <sample>.<haplotype>.yak.qv.txt
```

Assembly QV was also determined using Meryl and Merqury (Rhie et al., 2020). Meryl generates kmer databases and Merqury determines haplotype QV jointly with both haplotypes. Kmer databases with Meryl are generated with the following commands. Databases were generated separately for each read file using `meryl count` and merge with `meryl union-sum`. Parental-specific kmers (hapmers) were generated using `meryl hapmer`.

```
meryl k=21 threads=64 memory=32 count output <sample>.meryl
<read_file>
meryl union-sum output <sample>.meryl <sample_read_meryl_files>
bash hampers.sh maternal.meryl paternal.meryl sample.meryl
```

QV estimation with Merqury is generated with the following command:

```
merqury.sh sample.meryl maternal.meryl paternal.meryl
<maternal_haplotype> <paternal_haplotype> <sample>.merqury
```

GIAB-based Assembly Quality Assessment

As a complementary and stratified assessment of assembly quality, we used the GIAB assembly benchmarking pipeline to compare assembly-based variant calls to GIAB's v4.2.1 small variant benchmarks for two GIAB samples assembled in this work - HG002 and HG005. We evaluated the HG002 and HG005 HPRC year 1 assemblies aligned to GRCh38. Variants were called from assemblies using Dipcall v0.3 (using mimimap2 v2.2.4) (Li et al., 2018). We used `-z200000,10000` parameter to improve alignment contiguity, as previously shown to improve variant recall in regions with dense variation like the Major Histocompatibility Complex

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

(Chin et al., 2020). Small variant evaluation was performed using hap.py v3.15 (Krusche et al., 2019), benchmarking against v4.2.1 of high-confidence SNP, small indel, and homozygous reference calls for the GIAB samples HG002 and HG005. Comparisons were performed with and without restriction to the associated dipcall region file (dip.bed) to assess recall within and outside assembled regions. For better comparisons of complex variants, hap.py was run using vcfeval (Cleary et al., 2015). Variant calls were stratified using GIAB stratifiations v3.0 (J. Zook, 2021), stratifying true positive, false positive, and false negative variant calls in challenging and targeted regions of the the genome.

Trio Based Assembly Phasing Assessment

Assembly phasing was assessed using yak and is described using two statistics: switch error and Hamming error rates. Switch error describes the number of times two adjacent phased variants incorrectly switch between maternal and paternal haplotypes. Hamming error rate relates to the total number of misphased variants per assembled contig. Yak generates phasing statistics separately for each haplotype using parental kmers gathered from Illumina short-read sequencing of the parents.

Yak generates kmer databases for the sample and both parental haplotypes (as described above). Yak generates phasing metrics with the following command:

```
yak trioeval -t 32 paternal.yak maternal.yak <haplotype_assembly> >
<sample>.<haplotype>.yak_phasing.txt
```

Hi-C Based Assembly Phasing Assessment

An alternative approach for phasing evaluation is to use Hi-C reads that do not require trio information. We compute the switch error rate for local phasing evaluation and the hamming error rate for global phasing evaluation. We implement an efficient k-mer based method in `pstools-v0.1` (Garg, 2020) and use maximum Hi-C read support to detect switch errors on heterozygous positions. In this procedure, first, we find heterozygous k-mers (hets) from phased assemblies using 31-mers. Then, we map Hi-C reads to the assemblies using these 31-mers. If there are >5 reads that support a switch between consecutive hets in assemblies, we consider a haplotype switch. For each het pair, we note if Hi-C reads support or do not support the phase. We consider a switch error when a het site has phase switched support relative to that of the previous heterozygous site. The switch error rate is the number of local switches divided by the number of heterozygous sites. We perform this operation for the whole contig over all contigs for switch calculations. In the hamming error calculations, we consider hamming distance on the whole contig level divided by the number of heterozygous sites. This measure gives a global view of phasing errors and implicitly penalizes any long switches in contigs.

Flagger: Assembly Read-based evaluation

The following describes generating and cleaning the HiFi alignments to the HPRC/HPRC+ assemblies and running Flagger-v0.1, a read-based pipeline for evaluating diploid/dual assemblies. All the WDL-based workflows for running these steps are deposited in the

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

dockstore collection (https://dockstore.org/organizations/HumanPangenome/collections/Flagger-Secphase).

Preparing the HiFi alignments:

We aligned back the HiFi reads of each sample to its diploid assembly. The alignments were produced with `winnowmap-v2.03` using these commands:

```
 # making the k-mer table with meryl
  meryl count k=15 output merylDB asm.fa
  meryl print greater-than distinct=0.9998 merylDB > repetitive_k15.txt

  # alignment with winnowmap
  winnowmap -W repetitive_k15.txt -ax map-pb -Y -L --eqx --cs -I8g <(cat
pat_asm.fa mat_asm.fa) reads.fq.gz | samtools view -hb > read_alignment.bam
```

For all samples we used the full HiFi read sets mentioned in (**Supplementary Table 1**), except HG002 for which we downsampled the read set to 35X.

In order to exclude unreliable alignments, we removed all chimeric alignments and alignments shorter than 2kb or with a gap-compressed mismatch ratio higher than 1%. Since the assembly is diploid and the reads aligned to the homozygous regions are expected to have low mapping qualities, we didn't filter alignments based on their mapping qualities. In (**Supplementary Figure 44)**, we plotted the histograms of mapping qualities and the distributions of alignment identities for one sample, HG00438, as an example. The statistics of three sets of alignments are plotted; the alignments to the diploid assembly and to each haploid assembly (maternal and paternal) separately. It indicates that the reads have higher identities when the diploid assembly is used as reference but about 20% more reads have mapping qualities lower than 10.

Generally in highly homozygous regions, the aligner may not be able to select the correct haplotype as the primary alignment because of either read errors or misassemblies. To detect these cases we searched for secondary alignments whose scores are almost as high as the primary alignment of the same read. For each such read, we made a pseudo-multiple alignment of the read sequence and the assembly blocks captured by all secondary and primary alignments. Using this alignment, we searched for the read bases that are mismatched in at least one alignment but not all alignments. We called such bases single nucleotide markers. For each alignment we calculated a consistency score by considering only the single nucleotide markers and taking the summation of their base qualities with a negative sign. We then sorted the alignments (regardless of being primary or secondary) based on this score. If the best alignment was a secondary alignment we assigned the primary tag to this alignment and removed the other alignments. The percentage of the total reads with swapped alignments ranges from 0.03% (HG03453) to 0.44% (HG005) across 47 HPRC/HPRC+ samples. This shows that only a small percentage of the reads needed to be relocalized using this method. This step is performed through the Secphase-v0.1 workflow which is available in the dockstore

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

collection(https://dockstore.org/organizations/HumanPangenome/collections/Flagger-Secphase) .

By calling variants it is possible to detect the regions that either need polishing (i.e. are errors) or that have alignments from the wrong haplotype because of mismappings. We used DeepVariant `1.3.0` with the parameter *`--model_type="PACBIO"`* to call variants on these alignments. The variants were then filtered to include only the biallelic SNPs with variant frequency higher than 0.3 and genotype quality higher than 10.

```
bcftools view -Ov -f PASS -m2 -M2 -v snps -e 'FORMAT/VAF < 0.3 || FORMAT/GQ < 10' ${OUTPUT_VCF} > ${SNPS_VCF}
```

Having the biallelic SNPs we found the alignments with alternative alleles and removed them from the bam file. For this aim we implemented and used the program `filter_alt_reads`, running the command:

```
filter_alt_reads -i ${INPUT_BAM} -o ${ALT_FILTERED_BAM} -f ${ALT_BAM} -v ${SNPS_VCF}
```

Running the evaluation pipeline

To assess the read mappings resulting from our diploid alignment process we employed the following steps, which are combined into a pipeline that we refer to as Flagger. Flagger essentially fits a mixture model to successive coverage blocks of the read-to-diploid assembly alignment and then classifies each block to a category predicting the accuracy of the assembly at that location.

Step 1: Calculating depth of coverage

After producing and cleaning the HiFi alignments we calculated the depth of coverage for each assembly base by *`samtools depth -aa`*. (*`-aa`* option allows outputting the bases with zero coverage)

```
samtools depth -aa -Q 0 read_alignment.bam > read_alignment.depth
```

The output of samtools depth was then converted into a more efficient format with the `.cov` suffix. This format is implemented specifically for Flagger and it is more efficient since the consecutive bases with the same coverage take only one line. We implemented a program called `depth2cov` for converting the output of `samtools depth` to the `.cov` format.

```
depth2cov -d read_alignment.depth -f asm.fa.fai -o read_alignment.cov
```

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Step 2: Fitting the mixture model

The frequencies of coverages can be calculated with `cov2counts`. The output file with `.count` suffix is a 2-column tab-delimited file; the first column shows coverages and the second column shows the frequencies of those coverages.

```
cov2counts -i read_alignment.cov -o read_alignment.counts
```

The python script `fit_gmm.py` takes a file `.counts` suffix, fits a Gaussian mixture model, and finds the best parameters through Expectation-Maximization (EM). This mixture model consists of 4 main components and each component represents a specific type of region. The 4 components are

1. **Erroneous component**, which is modeled by a Poisson distribution. To avoid overfitting, this mode only uses the coverages below 10 so its mean is limited to be between 0 and 10. It represents the regions with very low read support.
2. **(Falsely) Duplicated component,** which is modeled by a Gaussian distribution whose mean is constrained to be half of the haploid component's mean. It should mainly represent the falsely duplicated regions.
3. **Haploid component,** which is modeled by a Gaussian distribution. It represents blocks with the coverages that we expect for the blocks of an error-free assembly.
4. **Collapsed component,** which is actually a set of components each of which follows a Gaussian distribution whose mean is a multiple of the haploid component's mean. It represents regions that have additional copies present in the underlying genome that have been "collapsed" into a single copy.

It was noticed that the model components may change for different regions due to regional coverage differences and that the resulting systematic differences affect the accuracy of the partitioning process. In order to make the coverage thresholds more sensitive to the local patterns the diploid assembly was split into windows of length (5-10Mb) and a distinct model was fit for each window. Before fitting, we split the whole-genome coverage file produced in step 1 into multiple coverage files for each window. We implemented and ran `split_cov_by_window` for splitting:

```
split_cov_by_window -c read_alignment.cov -f asm.fa.fai -s 5000000 -p
${OUTPUT_PREFIX}
```

`It will` produce a list of coverage files, each of which ends with `${CONTIG_NAME}_${WINDOW_START}_${WINDOW_END}.cov`

We then repeated steps above for each resulting coverage file.

One important observation is that for short contigs the coverage distribution is generally too noisy to satisfactorily fit the mixture model. To address this issue we have done the window-specific coverage analysis only for the contigs longer than 5Mb and for the shorter contigs we use the results of the whole-genome analysis.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Step 3: Extracting blocks of each component

Using the fitted model we assigned each coverage value to one of the four components (erroneous, duplicated, haploid, and collapsed). To do so for each coverage value we picked the component with the highest probability. For example, the coverage value 0 is frequently assigned to the erroneous component. In **(Supplementary Figure 45)** the coverage intervals are colored based on their assigned component.

Step 4: Incorporating coverage biases in HSats

According to the recent complete human genome paper (Nurk et al., 2022), there exist some satellite arrays (e.g. HSAT1,2,3) where the HiFi coverage drops or increases systematically due to biases in sample preparation and sequencing. Such platform-specific biases mislead the pipeline. As a result, the falsely duplicated component may contain a mixture of falsely duplicated and coverage-biased blocks. Similarly for the collapsed component.

To incorporate such coverage biases and correct the results in the corresponding regions, we first found the regions of each haploid assembly where a coverage bias is expected. To find such regions we lifted over the CHM13 HSat1, 2, and 3 annotation to each assembly by aligning the assembly contigs to the the reference T2T-CHM13v1.1 + GRCh38-chrY and projecting the HSat coordinates back to the assembly (using python script project_blocks.py). Then we ran fit_gmm.py to fit a mixture model for the blocks assigned to each HSat type and adjusted the parameter --coverage, the starting point of the EM process, based on the expected coverage in the corresponding HSat. For HSat1, 2 and 3 we set --coverage to 0.75, 1.25 and 1.25 times the average sequencing coverage respectively. Finally we decomposed each HSat based on the inferred coverage thresholds and replaced the previous assigned component by the new one.

Step 5. Using high quality alignments to correct spurious flags

In some cases the duplicated component is mixed up with the haploid one. It usually happens when the coverage in the haploid component drops systematically or the majority of a long contig is falsely duplicated. To address this issue we used another indicator of a false duplication, which is the accumulation of alignments with very low MAPQ. Therefore we have produced another coverage file using only the alignments with MAPQ > 20. Whenever we found a region flagged as duplicated with more than 5 high quality alignments we changed the flag to haploid.

After the correction made in step 5 we merged each components' blocks closer than 1k and the overlap of any two components after merging is flagged as Unknown to show that this block couldn't be assigned properly. The BED files produced by Flagger are available in the HPRC S3 bucket (https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=submissions/e9ad8022-1b30-11ec-ab04-0a13c5208311--COVERAGE_ANALYSIS_Y1_GENBANK/FLAGGER/APR_08_2022/FINAL_HIFI_BASED/FLAGGER_HIFI_ASM_SIMPLIFIED_BEDS/).

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Assessing T2T-CHM13 Using Flagger

To estimate the false positive rate of Flagger we applied it to the T2T-CHM13v1.1 reference.The direct output of Flagger showed that about ~12.77Mb (~0.41%) of the T2T-CHM13 reference assembly is flagged as potentially unreliable. The HPRC assemblies are almost free of rDNA arrays but there exist modeled sequences for rDNA arrays in the T2T-CHM13v1.1 reference. These arrays are flagged as falsely duplicated in their entirety indicating that Flagger with HiFi reads may not be able to evaluate rDNA arrays correctly. Therefore to make a fair comparison we excluded rDNA arrays (~9.92Mb in total) from the reference evaluation, which decreased the number of unreliable bases to 5.58Mb (~0.18%). We additionally identified ~2.76Mb of a region beside Chr1-HSat2 that was mis-flagged as collapsed. This mis-flagging was the impact of the systematic coverage rise in the neighboring HSat2 that altered the fitted mixture model. By manually fixing this mis-flagging we had ~2.82Mb (0.09%) of unreliable blocks in T2T-CHM13v1.1. This number is about 9.3 times lower than the average for the HPRC/HPRC+ assemblies. These unreliable blocks are mainly a combination of "Unknown" blocks which couldn't be assigned properly and also the regions with HiFi-specific coverage drops. The results of this analysis is available in the HPRC S3 bucket (https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=submissions/e9ad8022-1b30-11ec-ab04-0a13c5208311--COVERAGE_ANALYSIS_Y1_GENBANK/FLAGGER/APR_08_2022/FINAL_HIFI_BASED/T2T-CHM13/).

Repeat Masking

Repeat masking on each assembly was performed iteratively by RepeatMasker 4.1.2-p1. The first step masked using the default human repeat library, and the second step using a repeat library augmented by CHM13 satellite DNA sequences on the original assemblies after hard-masking the initial repeat masked DNA. The augmented repeat library is available at (https://zenodo.org/record/5537107#.YqNs13XMJrk), final_consensi_gap_nohsat_teucer.embl.txt, and a parallelized repeat masking pipeline is at (https://github.com/chaissonlab/segdupannotation), RepeatMaskGenome.snakefile. The union of the two steps yields the complete repeat masking.

Segmental Duplication Annotation

SDs were annotated using sedef (Numanagic et al., 2018) after masking repeats in each assembly. Repeats annotated with more than 20 copies were excluded from the analysis. The pipeline for annotating SDs is available at (https://github.com/ChaissonLab/SegDupAnnotation/releases/tag/vHPRC) .

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Unreliable Segmental Duplication Analysis

The reliable/unreliable regions for all haplotype assemblies were aligned to T2T-CHM13v2.0 and then subdivided into 5 kb windows and intersected with the SD annotations for T2T-CHM13. SD annotations were unavailable for chromosome Y on T2T-CHM13v2.0 at the time of analysis so chromosome Y was excluded. For each class of unreliable region (Unk,Err,Col,Hap) we calculated the average number of bp overlapping SDs across the haplotype assemblies, and annotated each 5 kb window with the most representative overlapping SD (the SD with the highest product of identity and length). Then using the most representative SD we calculated the average length and identity of SDs overlapping each class of unreliable region for all the 94 haplotypes and compared the length of identity of SDs that overlapped the different types of errors in the assembly. The code for this analysis is made available on github: (https://gist.github.com/mrvollger/3bdd2d34f312932c12917a4379a55973)

Ensembl Mapping Pipeline for Assembly Annotation

A reference gene set was created from a subset of the GENCODE 38 (Frankish et al., 2021) genes, which was mapped to the HPRC assemblies via a 2-pass alignment process. This excluded readthrough genes and genes on patches or haplotypes, and only included one copy of the genes on the X/Y PAR region (only one copy, the X, is modeled in the Ensembl representation of the PAR genes).

Firstly, in order to minimize the difficulty of mapping near identical paralogues, a jumping window of 100kb in length was used to identify clusters of genes to map in parallel (see **Supplementary Figure 46**). The initial window was positioned at the start of the most 5' gene for each chromosome in the GRCh38 reference and extended 100kb from the start of the gene. Any genes fully or partially overlapping the window were then included in the cluster. The next 3' gene that did not overlap the previous window was then identified and a new window was created and the process repeated. This resulted in both clustered genes and non-clustered genes (genes were considered not clustered when there was only one gene within the window). The regions to map were then identified based on the start of the most 5' gene and the end of the most 3' gene in each cluster (or simply the 5' and 3' end of the gene in the case of non-clustered genes).

For each region defined in the previous step, anchor points were then selected to help map the region on the target genome. Two 10kb anchor points were created 5kb from the 5' and 3' edge of the region, and a central 10kb anchor was created around the midpoint of the region in the GRCh38 genome. The sequences of these anchors were them mapped against the target genome with minimap2 (Li, 2018b) using the following command:

```
minimap2 --cs —secondary=yes -N 10 -x map-ont [genome_index] [anchor_file] >
[alignment_file]
```

The resulting hits were examined to determine high confidence regions in the target genome. High-confidence regions were ones in which all three anchors were on the same top-level

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

sequence, in co-linear order, with >= 99 percent sequence identity and >= 50 percent hit coverage, and with a similar distance between the anchors when compared to the reference genome. If no suitable candidate region was found with all three anchors, pairs of mapped anchors were then assessed in a similar manner.

The sequence selected region or regions were then retrieved and aligned against the corresponding GRCh38 region using MAFFT. For each gene, the corresponding exons were retrieved and the coordinates were projected through the alignment of the two regions. Transcripts were then reconstructed from the projected exons. For each transcript, the coverage and identity when aligned to the parent transcript from GRCh38 was calculated.

If the resulting transcript had either a coverage or < 98 percent or identity of < 99 percent, the parent transcripts were aligned to the target region using minimap2 in splice-aware mode, with the high quality setting for Iso-Seq/cDNA style transcripts enabled. The maximum intron size was set to 100kb by default. For transcripts with reference introns larger than 100kb, the max intron size was scaled and set as 1.5 times the length of the longest intron (to allow some variability).

```
minimap2 --cs --secondary=no -G [max_intron_size] -ax splice:hq -u b
[expected_target_region] [transcript_sequences] > [sam_file]
```

For each transcript that mapped to the target genome, the quality of the mapping was assessed based on aligning the original reference sequence with the newly identified target sequence. Again, if the coverage or identity of the aligned sequence was < 98/99 percent, the reference transcript sequence was re-aligned to the target region, this time using Exonerate (Slater & Birney, 2005). Exonerate, while slower than minimap2, has the ability to handle very small exons and also can incorporate CDS data to preserve the CDS (introducing pseudo-introns as needed). The following command was used:

```
exonerate -options --model cdna2genome --forwardcoordinates FALSE --
softmasktarget TRUE --exhaustive FALSE --score 500 --saturatethreshold 100 --
dnawordlen 15 --codonwordlen 15 --dnahspthreshold 60 --bestn 1 --maxintron
[max_intron_size] -coverage_by_aligned 1 --querytype dna --targettype
[target_type] --query [query_file] --target [target_file] --annotation
[annotation_file] > [output_file]
```

When more than one approach was used to model the transcript, the mapping with the highest combined identity and coverage was selected.

For genes not mapped through the initial regional anchors, a second approach was used. The expected location of the gene was located using high confidence genes mapped during the first phase. High confidence mappings were those which the there was a single mapped copy of the gene, all the transcripts had mapping scores of 99 percent coverage and identity on average and the gene also had a similar gene neighborhood to the neighborhood in the reference (at

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

least 80 percent of the of the same genes in common for the 100 closest neighboring genes in the reference). After this, the entire genome region underlying the missing gene, including 5kb flanking sequence, was mapped it against the target genome using minimap2:

```
minimap2 --cs —secondary=yes -x map-ont [genome_index]
[gene_genomic_sequence] > [alignment_file]
```

The resulting hits were then filtered based on overlap with the expected region the missing gene should lie in. If there was no expected region calculated (cases where no pair of high confidence genes could be found to define the 5' and 3' boundaries of the expected location of the missing gene, e.g. at the edge of a scaffold), or no hit overlapping the expected region was found, the top reported hit was used providing it passed an identity cut-off of 99 percent. The selected hit or hits were then extended based on how much of the original reference gene they covered, to ensure minor local variants between the reference and target regions did not lead to the target region being truncated. Once extended, the remaining hits were then clustered based on genomic overlap and merged into unique regions. The missing genes were then attempted to be mapped to these regions using an identical process as described above for the initial mappings, involving MAFFT, minimap2 and Exonerate.

In order to try and minimize the occurrences of mismapped paralogues, each gene was checked in terms of the genes with exon overlap in both the target and the reference. If the overlapping genes were not identical at a locus between the reference and the target then a conflict was identified. For each gene present, filtering was done to reduce or remove the conflict based on a number of factors including whether the genes were in the expected location, whether the genes were high confidence mappings, the average percent identity and coverage of the transcript for the genes and the neighborhood score. When it was not possible to resolve a conflict between two genes, both were kept. This concluded the primary mapping process.

After this potential recent duplications were identified. To search for recent duplications, the canonical transcript of each gene (the longest transcript in the case of non-coding genes, or the transcript with the longest translation followed by the longest overall sequence for protein-coding genes) was selected, and aligned across the genome using minimap2 in a splice-aware manner:

```
minimap2 --cs --secondary=no -G [max_intron_size] -ax splice:hq -u b
[genome_index] [input_file] > [sam_file]
```

Mappings that had exon overlap with existing annotations from the primary mapping process on the target genome were removed. For new mappings that did not overlap existing annotations, the quality of the alignment was then assessed by aligning the mapped transcript sequence to the corresponding reference transcript to calculate the coverage and percent identity of the mapping. Different coverage and percent identity cutoffs were used for these mappings, based on the type of transcript mapped. The cutoffs for retaining were as follows:

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Transcript type	Coverage cutoff (%)	Identity cutoff (%)
Protein coding	95	95
Long non-coding	90	90
Small non-coding	95	95
Pseudogene	80	90

When looking for new paralogues, in case that multiple canonical transcripts mapped to a locus, a single representative transcript was selected. This was based on the following heirarchy of gene biotype groups:

- Coding
- Long non-coding
- Pseudogene
- Small non-coding
- Misc/Undefined

If there were multiple transcripts for the highest represented group, the transcript with the longest sequence was selected as the representative.

Gene annotation quality analysis

Frameshifts

For the Ensembl and CAT gene annotation sets, we identified the locations of frameshifting insertions and deletions by iterating over the coding sequence of each transcript and looking for any gaps in the alignment. If the gap had a length that was not a multiple of 3, and its length was less than 30 base pairs long (to remove likely introns from consideration), the gap is determined to be a frameshift and its location is saved to a BED file.

Nonsense mutations

We also analyzed the number of nonsense mutations that would cause early stop codons in both the Ensembl and CAT gene annotation sets. We identified the nonsense mutations by iterating through each codon in the coding sequence of the predicted transcripts, and if there was an early stop codon before the canonical stop codon at the end of the transcript, we saved the location in a BED file.

Validation of mutations with Illumina

For both sets of mutations, we then lifted over the coordinates of the mutations to be on the GRCh38 reference so that we could use existing variant callsets on GRCh38. We used `halLiftover` to lift over each set of coordinates, using the GRCh38-based HAL file from the

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

cactus-minigraph alignment. Then, we used `bedtools intersect` to intersect with the variant call file for each of the assemblies.

Sample commands:

```
halLiftover GRCh38-f1g-90-mc-aug11.hal <GENOME_NAME>
<MUTATION_BED_FILE> GRCh38 <LIFTED_OVER_BED_FILE>

bedtools intersect -wo -a <LIFTED_OVER_BED_FILE> -b
<SAMPLE_MERGED_VCF>  > <OVERLAP_OUTPUT_TXT_FILE>
```

The VCF files used in this intersection were downloaded from the 1KG: (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20201028_3202_raw_GT_with_annot/20201028_CCDG_14151_B01_GRM_WGS_2020-08-05_chr$i.recalibrated_variants.vcf.gz)
Where `$i` was replaced with each chromosome number. From there, each chromosome vcf was split so that each sample was in its own file using `bcftools view`. The chromosome files for each sample were combined into one VCF using `bcftools concat`.

Gene Duplication Analysis

Duplicated genes were detected as multi-mapped coding sequences using Liftoff (Shumate & Salzberg, 2020) supplemented by a complementary approach (gb-map) using multi-mapped gene bodies. The combined set was formed by including all liftoff gene duplications and duplicated genes detected by gb-map.

Liftoff

We ran Liftoff (commit `35a4e5536414c4ac3b49873f427388d54bc24fd7`) to annotate extra gene copies in each of the assemblies. Liftoff was run with the flag `-sc = 0.90` to find additional copies of genes, with an identity threshold of at least 90%. An example command is below:

```
liftoff -p 10 -sc 0.90 -copies -db <GENCODE_V38_DATABASE> -u <UNMAPPED_FILE>
-o <OUTPUT_GFF3> -polish <GENOME_FASTA> <GRCh38_FASTA>
```

The additional copies of the genes are identified as such in the output gff3 with the field `extra_copy_number` (equal to anything other than 0). For this analysis, we also only considered genes that were multi-exon, protein-coding genes. The additional gene copies were further filtered to remove any genes outside of the “reliable”, haploid regions as determined by the Flagger pipeline.

gb-map

The gene-body mapping pipeline identifies duplicated genes by first aligning transcripts of protein coding and pseudogenes (GENCODE v38) to each assembly, and then multi-mapping the genomic sequences of each corresponding gene. Alignments of at least 90% identity and 90% of the length of the original duplication are considered candidate duplicated genes.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Candidates are removed if they overlap previously mapped transcripts from other genes, low-quality duplications, and genes identified through CAT and liftoff analysis.

Gene family analysis

To account for gene duplications in high-identity gene families, gene families are identified based on sequence alignments from gb-map. Genes that map reciprocally with 90% identity and 90% length are considered a gene family. A single gene is selected as the representative gene for the family, and any gene duplication in the family is counted towards that gene.

Pangenome Graph Construction

Minigraph

Minigraph can quickly perform assembly-to-graph mappings using a generalization of the minimap2 algorithm (Li et al., 2020). Novel SVs of at least 50 bp detected in the mapping can then be added to the graph. To construct a pangenome graph, one chosen reference assembly, GRCh38 in this case, is used as a starting graph, and the mapping and SV addition steps are repeated for each additional assembly, greedily. This iterative approach is analogous to Partial Order Alignment (POA) (Lee et al., 2002). Graphs constructed in this way describe the structural variation within the samples and provide a coordinate system across the reference and all insertions. Minigraph does not produce self-alignments. That is, it will never align a portion of the reference assembly onto another portion of the reference assembly. In this way all reference positions have a unique location within the created pangenome. Minigraph version v0.14 was used with `-xggs` options. The input order was GRCh38, CHM13 then the remainder in lexicographic order by sample name.

Minigraph-Cactus

Graphs constructed by Minigraph only contain structural variation (≥50 bp) by default. The aim of the MC pipeline is to refine Minigraph's output in order to include smaller variants, down to the SNP level. Doing so allows the graph to comprehensively represent most variation, as well as to embed the input haplotypes within it as paths, which is important for some applications (Sirén et al., 2021). This pipeline is composed of the five following steps, and is described in more detail in (Hickey et al., in preparation). The script and commands to reproduce can be found at (https://github.com/ComparativeGenomicsToolkit/cactus/blob/81903cb82ae80da342515109cdee5a85b2fde625/doc/pangenome.md#hprc-version-10-graphs). A newer, simpler version of the pipeline that no longer requires satellite masking can be found at (https://github.com/ComparativeGenomicsToolkit/cactus/blob/5fed950471f04e9892bb90531e8f63be911857e1/doc/pangenome.md#hprc-graph).

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Paths from the reference, GRCh38, are acyclic in the MC graph. Paths from any other haplotypes can contain cycles (as a result of different query segments mapping to the same target), but they are relatively rare.

1. **Satellite Masking**: Minigraph is unable to map through highly repetitive sequence such as centromeres and telomeres and, since these regions are also enriched for misassemblies (see Assembly Assessment subsection of Results), we decided to explicitly exclude them from the MC graphs used in this work. dna-brnn is a tool that uses a Recurrent Neural Network (RNN) to quickly identify alpha satellite as well as human satellite I and II (Li, 2019c). We ran it with its default parameters on all input sequences and cut out any identified regions >= 100kb, except on the reference. The three satellite families that dna-brnn detects account for the vast majority of satellite sequence, but not all. As such, gaps >=100kb in minigraph mappings were also removed. They were detected by mapping each assembly, after having removed the dna-brnn regions, to the minigraph (using the procedure described below). In all, an average of 188.6mb of sequence from each (non-reference) assembly was excluded from the graph.
2. **Assembly-to-Graph Mapping**: Minigraph generalizes minimap2's fast seeding and chaining algorithms, but it does not currently produce exact alignments in cigar strings or otherwise. For this work, an option, `-write-mz`, was added to report chains of minimizers, which in this case are 15bp exact matches, and all assemblies were mapped to the minigraph graph using it. The resulting minimizers were then converted into PAF files with cigars representing exact pairwise alignments between the query contigs and minigraph node sequences, and all mappings with MAPQ<5 were excluded, as were overlapping *query* regions>10kb.
3. **Chromosome Decomposition**: The Minigraph graphs do not contain inter-chromosomal rearrangements, but the mappings performed in the previous steps can imply them: i.e. a contig can partially map to multiple chromosomes. In most cases, these mappings involve similarity across different acrocentric short arms. In order to avoid introducing misleading interchromosomal events, and because it is necessary to run the subsequent steps individually on chromosomes due to memory requirements, the mappings were divided by reference chromosome. This was done by splitting the minigraph into connected components and using the RN tags to determine their corresponding chromosome names. The PAF mappings were used to determine the coverage of each query contig with each chromosome component. This coverage was used to assign each query contig to a single chromosome by choosing the chromosome with the highest coverage. Contigs with insufficient coverage to any chromosome (<90% for contigs with lengths in [1,10kb); <80% for [10kb,100kb), <75% for [100kb,1mb) and <70% for >=mb.) were considered ambiguous and not included in the graph.. In the GRCh38-based graph, all unplaced and random contigs were grouped together into the same component.
4. **Cactus Base Alignment**: Cactus is a tool that uses a graph-based approach to combine sets of pairwise alignments obtained from lastz into a multiple genome alignment (Armstrong et al., 2020). When aligning different species, it uses a phylogenetic tree to

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

progressively decompose the alignment into a subproblem for each ancestral node in the tree. We adapted it to also accept chromosome-scale sequence-to-minigraph mappings as produced above, and improved its runtime on alignments of many sequences by replacing its base aligner with abPOA (Gao et al., 2021). The core algorithm described in (Armstrong et al., 2020) remains unchanged: the pairwise alignments are used to induce a sequence graph, then filtered using the Cactus Alignment Filtering (CAF) algorithm, and components of unaligned sequence are then processed by the Base Alignment and Refinement (BAR) algorithm. The resulting graph is used to infer an ancestral sequence (not explicitly used in this work), and then exported to a Hierarchical ALignment (HAL) file (Hickey et al., 2013). We implemented a converter, hal2vg (Hickey, 2021) that converts the HAL alignment into a sequence graph in vg format.

5. **Post-processing and Whole-Genome Indexing**: The following post-processing steps were performed on each chromosome graph. First, unaligned sequence >10kb in length, including sequence not aligned to minigraph, was removed in order to filter out any under-alignment artifacts that might later be mistaken for insertions. Next, GFAffix was used to normalize the graphs by merging together redundant node prefixes and suffixes. Nodes were flipped as necessary to ensure that reference paths always visit their forward orientations. The chromosomes were combined into a whole-genome graph, indexed and exported to VCF, all using vg. Patched versions of both the GRCh38- and CHM13-based graphs were created when it was discovered that short contigs split-mapping to distant locations had induced large deletions. The deletions were removed using vg clip -D 10000000 (and the pipeline has since been corrected to no longer produce them). Allele-filtered graphs, used for short-read mapping, were produced (from the patched graphs) by removing all nodes traversed by fewer than 9 haplotype paths (minimum AF=10%) using vg clip -d 9 -m 10000. The chromosome HAL files were also combined into a whole-genome HAL file using halMergeChroms, and clipped sequences added back (in order to facilitate running CAT) using halUnclip.

PanGenome Graph Builder (PGGB)

The Pangenome Graph Builder (PGGB) uses a symmetric, all-to-all comparison of genomes to generate and refine a pangenome. We applied it to build a pangenome graph from all genome assemblies and references (both GRCh38 and CHM13) in the HPRC year-1 freeze. The resulting PGGB graph represents all alignment relationships between input genomes in a single graph. The PGGB graph is a lossless model of the input assemblies that represents all equivalently. This arrangement allows for all of our pangenome assemblies to be used as reference systems, a property that we used to explore the scope of pangenome variation in a total way. We apply the PGGB model to investigate the full pangenome and integrate annotations established *de novo* on the diverse assemblies into a single model for analyses of pangenome diversity, and also of complex structurally-variable loci (MHC and 8p inversion).

PGGB generates a pangenome graph in three phases. (1) *alignment*: In the first, phase the wfmash aligner is used to generate all-vs-all alignments of input sequences. This method, wfmash, applies MashMap2's mapping algorithm to find homologies at a specified length and

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

percent identity. It then derives base-level alignments using a high-order version of the WFA algorithm (wflign) which first aligns sequences in segments of 256bp, then patching up the base-level alignment with local application of WFA. wfmash was designed and developed specifically for the problem of building all-to-all alignments for large pangenomes. (2) *graph induction*: The input FASTA sequences and PAF-format alignments produced by wfmash are converted to a graph (in GFA format) using seqwish. This losslessly transforms the input alignments and sequences into a graph. (3) *graph normalization*: We finally apply a normalization algorithm—smoothxg—to simplify complex motifs that occur in STRs and other repetitive sequences, as well as mitigate underalignment. The graph is first sorted using a "path-guided" stochastic gradient descent method (Guarracino, Heumos, et al., 2022) that organizes the graph in 1 dimension so as to optimize path distances and graph distances. This sort provides a way to partition the graph into smaller pieces over which we apply a multiple sequence alignment algorithm (abPOA). These pieces are laced back into a final graph. We iterate this process twice using different target POA lengths to remove boundary effects caused at the borders of the MSA problems. Finally, we apply GFAffix to remove redundant furcations from the topology of the graph.

To build the HPRCy1 PGGB graph, we used both the CHM13 and GRCh38 references as a target and mapped all contigs against these with wfmash requiring a full length mapping at 90% total identity, collecting all contigs that mapped to a given chromosome. Contigs which did not map under this arrangement were then partitioned using a split mapping approach, requiring 90% identity over 50kb to seed the mappings, and putting the contig into the chromosome bin for which it had the best split mapping. We thus initially partition the data into 25 chromosome sets: one for each autosome, one for each sex chromosome, and finally the mitochondria.

We then applied PGGB (version 0.2.0+531f85f) to each partition to build a chromosome specific graph. Run in parallel over 6 PowerEdge R6515 AMD EPYC 7402P 24-core nodes with 384GB of RAM, this process requires 22.49 system days, or around 3.7 days wallclock. (To develop a robust process to build the HPRCy1 graph, the pggb team iterated the build 88 times.) The final chromosome graphs were compacted into a single ID space using `vg ids -j`, then for each reference (GRCh38 and CHM13) a combined VCF file was generated from the graph with vg deconstruct (version 1.36.0 / commit 375cad7).

A handful of key parameters define the shape of the resulting graph. First, in wfmash we require >100kb mappings at 98% identity. We map each contig to all the other 89 input haplotypes. To reduce complexity, as well as false positive SNPs resulting from misaligned regions, we apply a minimium match length filter (in seqwish) of 311bp. This means that the graph which we induce is rather "underaligned" locally, and only via normalization in smoothxg do we compress the bubble structures that are produced. For smoothxg, our first iteration attempts to generate 13033bp-long POA problems, while the second is 13177bp. These lengths provided a balanced tradeoff between runtime and variant detection accuracy.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

In addition to a graph (in GFA), PGGB generates visualizations of the graph in 1D and 2D which show both the topology (2D) and path-to-graph relationship (1D). A code-level description of the build process is provided at (https://github.com/pangenome/HPRCyear1v2genbank).

Pangenome Graph Assessment

Annotating variant sites in pangenome graphs

Variant sites in Minigraph and MC/PGGB graphs were discovered using `gfatools bubble` (v0.5, (Li, 2021a)) and `vg deconstruct` (Paten et al., 2018), respectively. Large (>10 Mb) spurious deletions in MC/PGGB graphs were removed using vcfbub (v0.1.0, (Garrison, 2021)) with options `-l 0 -r 10000000`. Next, variant sites were classified into small variant (<50 bp) and SV (≥50 bp) sites. The SV sites were then annotated as described in the Methods section of minigraph paper (Li et al., 2020). In brief, the longest allele sequence of each SV site was extracted and stored in the FASTA format. The interspersed repeats, low-complexity regions (LCRs), exact tandem repeats, centromeric satellites, and gaps in the longest allele sequences were then identified using RepeatMasker (v4.1.2-p1) with NCBI/RMBLAST (v2.10.0) search engine and Dfam (v3.3) database, SDUST (v0.1, (Li, 2019b)), ETRF (Li, 2019a), dna-brnn (v0.1, (Li, 2019c)), and `seqtk gap` (v1.3, (Li, 2018a)), respectively. SDs were identified if total node length in a site is ≥1000bp and ≥20% of bases of these nodes are annotated as SD in the reference or in individual assembly (Methods). To find hits to the GRCh38 reference genome, minimap2 (v2.24) with options `-cxasm20 -r2k --cs` was used to align the longest allele sequences to the reference genome. Based on the identified features, SV sites were classified into various repeat classes using `mgutils.js anno` (https://github.com/lh3/minigraph/blob/master/misc/mgutils.js) with minor modifications to enable it to work with the files derived from MC and PGGB graphs.

Pangenome Size and Growth

We use the “heaps” tool of the odgi pangenome analysis toolkit (Guarracino, Heumos, et al., 2022) to estimate how the euchromatic autosomal pangenome grows with each additional genome assembly added. Here we approximate euchromatic regions by non-satellite DNA, which has been identified by dna-brnn in the construction of the MC graph (cf. Section **Minigraph-Cactus**) While the MC non-reference haplotypes of the MC graph do not contain satellite DNA, the PGGB graph does. Consequently, we subset the PGGB graph to segments contained in the MC graph. We additionally exclude reference haplotypes (GRCh38, CHM13) from the analysis. We then sampled permutations of the 88 non-reference haplotypes. In each permutation, we calculate the size of the pangenome after adding the first 1, 2, …, N haplotypes in both graphs. This yields a collection of saturation curves from which we derived a median saturation curve onto which we fitted a power law function known as Heaps’ Law. The exponent of this function is generally understood to represent the degree of openness—or diversity—of a pangenome (“Comparative Genomics: The Bacterial Pan-Genome,” 2008). Summing up, we

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

called `odgi heaps -i <graphs.gfa> -S -n200` to generate pangenome saturation curves for 200 permutations.
Next to calculating a non-permuted cumulative base count, we also counted the number of common (≥ 5% of all non-reference haplotypes) and core (≥ 95% of all non-reference haplotypes) bases in the pangenome graphs. To this end, we used a tool called "pangenome-growth" (Doerr, 2022) and supplied a list of the samples in which they are grouped according to their assigned superpopulation (`pangenome-growth -m -t bp <graph.gfa> <sample order>`). We repeated the count, this time including only segments of depth ≥ 2, i.e., contained at least twice in any haplotype sequence.

Decomposing pangenome graphs based on allele traversals

Pangenome graphs were decomposed topologically into a set of nested subgraphs, termed snarls, that each correspond to one or a collection of genetic variants. These snarls were then converted to VCF format using `vg deconstruct` (Paten et al., 2018). Large (>100 kb) deletions in MC/PGGB graphs were removed using vcfbub (v0.1.0, (Garrison, 2021)) with options `-r 100000`. To ease the comparison of variants with other call sets for each individual, the multi-sample VCF files were converted to per-sample VCF files using `bcftools view -a -I -s <sample name>` and the multiallelic sites were splitted into biallelic records using `bcftools norm -m -any`. Due to the limitations of snarl decomposition, snarls may contain multiple variants that cannot be further decomposed into nested snarls using `vg deconstruct`. If snarls of this kind are compared with truth calls, the evaluation will not be accurate. We solved this problem by comparing reference and alternate allele traversals for each snarl to infer the minimalist representation of variants (**Supplementary Figure 47)**.

Annotating small variants in pangenome graphs with allele frequencies from gnomAD

Variant sites in MC/PGGB graphs were discovered using `vg deconstruct` (Paten et al., 2018). The resulting VCF files were then decomposed based on allele traversals (Methods). The multi-nucleotide polymorphisms and complex indels were further decomposed into SNPs and simple indels using `vcfdecompose --break-mnps --break-indels` from RTG Tools (v3.12.1), (Cleary et al., 2015), so that they can be annotated with gnomAD later. For comparison, variants called from PacBio HiFi reads using DeepVariant and from haplotype-resolved assemblies using Dipcall were also used. For each discovery method, small variants (<50 bp) were extracted and normalized using `bcftools norm -c s -f <reference sequence in FASTA format> -m -any`. Next, all per-sample VCF files were combined into one VCF file using `bcftools concat -a -D` after dropping individual genotype information using `bcftools view -G`. To annotate small variants with allele frequencies from gnomAD (Karczewski et al., 2020), the gnomAD v3.1.2 per-chromosome VCF files were downloaded and concatenated into one VCF file using `bcftools concat`. The VCF file was then compressed into a file in the gnotate format using `make-gnotate` from `slivar` (v0.2.7, (Pedersen et al., 2021)) with options `--field AC:gnomad_ac --field AN:gnomad_an --field`

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

`AF:gnomad_af --field nhomalt:gnomad_nhomalt`. The small variants were annotated with gnomAD using `slivar expr --gnotate <gnotate file>`.

Variant benchmarking

Calling variants from PacBio HiFi reads

The PacBio HiFi reads were aligned to the GRCh38 human reference genome with no alternates using Winnowmap2 (v2.03, (Jain et al., 2022)) with `-x map-pb -a -Y -L --eqx --cs`. The MD tags required by Sniffles were calculated using `samtools calmd`. The resulting BAM files were sorted and indexed using SAMtools.

For small variants, the two-pass mode of DeepVariant (v1.1.0, (Jain et al., 2022)) with WhatsHap (v1.1, (Martin et al., 2016)) was used to call SNPs and indels from the PacBio HiFi read alignments. The resulting VCF files were used as truth sets for small variant benchmarking.

Three discovery methods were used to call SVs from the PacBio HiFi read alignments. For PBSV (v2.6.2, (Pacific Biosciences, 2021)), SV signatures were identified using `pbsv discover` with `--tandem-repeats <GRCh38 TRF BED file>` to improve the calling performance in repetitive regions. SV were then detected using `pbsv call` with `--ccs --preserve-non-acgt -t DEL,INS,INV,DUP,BND -m 40` from the signatures. For SVIM (v2.0.0, (Heller & Vingron, 2019)), SV were called using `svim alignment` with `--read_names --zmws --interspersed_duplications_as_insertions --cluster_max_distance 0.5 --minimum_depth 4 --min_sv_size 40`. In contrast to PBSV and Sniffles, SVIM outputs all calls no matter their quality. To determine the threshold used for filtering low-quality calls, a precision-recall curve was generated across various quality scores by comparing with the GIAB v0.6 Tier 1 SV benchmark set for HG002 (**Supplementary Figure 48**). Consequently, SVIM calls with a quality score lower than 10 were excluded. For Sniffles (v1.0.12b, (Sedlazeck et al., 2018)), SV were discovered with `-s 4 -l 40 -n -1 --cluster --ccs_reads`. Unlike PBSV and SVIM, Sniffles doesn't generate consensus sequences of insertions from aggregating multiple supporting reads. Therefore, Iris (v1.0.4, (Kirsche et al., 2021)) was used to refine the breakpoints and insertion sequences with `--hifi --also_deletions --rerunracon --keep_long_variants`. All resulting VCF files were sorted and indexed using BCFtools.

Calling SVs from haplotype-resolved assemblies

Three discovery methods were used to call SVs from the haplotype-resolved assemblies generated by Hifiasm.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

For SVIM-asm (v1.0.2, (Heller & Vingron, 2020)), assemblies were aligned to the GRCh38 human reference genome with no alternates using minimap2 (v2.21, (Li, 2018b)) with `-x asm5 -a --eqx --cs` and then sorted and indexed using SAMtools. SV were called using `svim-asm diploid` with `--query_names --interspersed_duplications_as_insertions --min_sv_size 40`. The resulting VCF files were sorted and indexed using BCFtools.

For PAV (v0.9.1, (Ebert et al., 2021)), assemblies were aligned to the GRCh38 human reference genome with no alternates using minimap2 (v2.21, (Li, 2018b)) with options `-x asm20 -m 10000 -z 10000,50 -r 50000 --end-bonus=100 --secondary=no -a --eqx -Y -O 5,56 -E 4,1 -B 5`. These alignments are then trimmed to reduce redundancy of records and increase contiguity of alignments. SVs, indels, and single-nucleotide variants were called by using cigar string parsing of the trimmed alignments. Inversion calling in PAV uses a novel k-mer density assessment to resolve inner and outer breakpoints of flanking repeats, which does not rely on alignment breaks to identify inversion sites. This is designed to overcome limitations in alignment methodologies and expand inversion calls which result in duplications and deletions of sequence on the boundaries.

The Hall-lab pipeline is as documented in the WDL workflow (https://github.com/hall-lab/competitive-alignment/blob/master/call_assembly_variants.wdl) (commit 0acce55). Briefly, the maternal and paternal assemblies were aligned to the GRCh38 human reference genome using minimap2 (v2.1 (Li, 2018b)) with options `-ax asm5 -L --cs`. Large indels (>50 bp) were detected using the 'call_small_variants' task, based on paftools (v2.17-r949-dirty). For large SV, breakpoints were mapped based on split alignments of assembly contigs to the reference genome and classified as SVs using a series of custom python scripts in the 'call_sv' task. The breakpoint-mapped SVs were then filtered based on the coverage of the reference genome by the assembly contigs (calculated using `bedtools genomecov`, v2.28.0). For each haplotype assembly, a BED file of 'exclude regions' was defined comprising genomic regions covered by more than one distinct contig or with more than 3X coverage by a single contig. Breakpoint-mapped SV where either breakpoint or >50% of the outer span intersected an exclude region were filtered.

Merging SV call sets

To integrate per-sample VCF files generated by three HiFi-based and three assembly-based SV callers, svtools (Larson et al., 2019) was used. For each individual, VCF files from the 6 callers were jointly sorted and then merged using svtools `lsort` and `lmerge`, first using a strict criterion (`svtools lmerge -f 20`), followed by a more lenient second merge (`svtools lmerge -f 100 -w carrier_wt`). The autosomal SV calls supported by at least two callers were included in the consensus SV call set for comparison.

Defining confident regions for variant benchmarking

For SVs, confident regions were generated using Dipcall. While useful for small variants, current benchmarking tools like hap.py/vcfeval cannot properly compare different representations of

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

small variants in and around SVs. Therefore, for each sample, the confident regions from Dipcall were further processed as follows:

1. Exclude any SD, self-chain, tandem repeat longer than 10 kb, or satellite DNA, if there are any breaks in the Dipcall BED file in the repeat region +15 kb flanking sequence on each side. The rationale is that breaks in the Dipcall BED file are generally caused by missing sequence or errors in the assembly or reference, or by large SVs or CNVs where we do not have tools to benchmark small variants in these regions
2. Exclude 15 kb around all breaks in the Dipcall BED file for the same reason as previous
3. Exclude 15 kb around all gaps in GRCh38 because alignments are unreliable
4. Exclude variants >49 bp in the Dipcall VCF file and any tandem repeats overlapping SVs +50 bp on each side

Benchmarking variants

Variant sites in MC/PGGB graphs were discovered using `vg deconstruct` (Paten et al., 2018). Variant sites with alleles larger than 100kb in MC/PGGB graphs were then removed using vcfbub (v0.1.0, (Garrison, 2021)) with options `-l 0 -a 100000`. The resulting VCF files were further processed using vcfwave from vcflib (Garrison, Kronenberg, et al., 2022) with option `-I 1000`. In brief, vcfwave realigned alternate alleles against the reference allele for each variant site using the bidirectional wavefront alignment (BiWFA) algorithm (Marco-Sola et al., 2022) to decompose complex alleles into primitive ones. The multi-sample VCF files were then converted to per-sample VCF files using `bcftools view -a -I -s <sample name>` and the multiallelic sites were splitted into biallelic records using `bcftools norm -m -any`. Next, the autosomal small variants (<50 bp) from a given pangenome graph (query set) were compared to the HiFi-DeepVariant call set (truth set) using vcfeval from RTG Tools (v3.12.1, (Cleary et al., 2015)) with options `-m annotate --all-records --ref-overlap --no-roc`. Note that the multi-nucleotide polymorphisms and complex indels were reduced to SNPs and simple indels using `vcfdecompose --break-mnps --break-indels` from RTG Tools (v3.12.1), (Cleary et al., 2015)The comparison was performed independently for each individual. Recall and precision were calculated within the refined Dipcall confident regions (Methods) and then stratified by the GIAB v3.0 genomic context. To evaluate the SV (≥50 bp) calling performance, the autosomal SVs from a given pangenome graph (query set) were compared to the consensus SV call set (truth set) for each individual using `truvari bench` (v3.2.0, (English et al., 2022)) with options `--multimatch -r 1000 -C 1000 -O 0.0 -p 0.0 -P 0.3 -s 50 -S 15 --sizemax 100000 --includebed <Dipcall confident regions>`. Recall and precision were then stratified by the GIAB v3 genomic context and by variant length.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Alignment of long-reads to pangenome graphs

PacBio HiFi reads

PacBio HiFi reads from 44 HPRC/HPRC+ samples were aligned to the Minigraph-Cactus graph using GraphAligner (v1.0.13, (Rautiainen & Marschall, 2020)) with option `-x vg` and stored in the GAF format (Li et al., 2020). A read might align to multiple places in the graph, the one with highest alignment score was kept and those with lower alignment score were dropped. To further remove low-quality alignments, a read with <80% of read length aligned to the graph was discarded. After filtering the read-to-graph alignments, the read depth of each edge was calculated using `vg pack` (v1.33.0, (Hickey et al., 2020)) with options `-Q -1 -D`. Note that the resulting GAF files didn't contain a mapping quality (encoded as 255 for missing) for each alignment, therefore the option `-Q -1` was given to `vg pack` to ensure that these alignments were used during read depth calculation. Next, the edges of each sample were classified into either on-target or off-target depending on whether they are on the sample paths (encoded as W-lines in Minigraph-Cactus GFA files) or not.

Oxford nanopore reads

ONT reads obtained from 29 HPRC/HPRC+ samples were aligned against the Minigraph-Cactus graph. The alignments were produced using GraphAligner v.1.0.13 with parameter settings "`-x vg --multimap-score-fraction 1 --multiseed-DP 1`". The number of reads in these data sets range between 1M and 5.4M and have an average read length of 28.4kb.
On average 99.68% of the reads received hits from one or more locations in the graph. For each read, we determined its best hit based on *alignment score* and discarded all its lower-scoring alignments in subsequent analysis. The alignment identities of these best hits peak well above 95% with an average ratio of *alignment-length-to-read-length* (ALRL) of 0.880 (std 0.302) and average MAPQ value of 59.35. The alignment set was further quality-pruned by discarding alignments that either had an ALRL lower than 0.8 or a MAPQ value lower than 50. The surviving alignments have an overall average ALRL of 0.968 (std 0.047) and effectuate an overall genome coverage between 10.5- and 43-fold across the 29 samples (**Supplementary Figure 49**).

Annotating genes within pangenome graphs

We ran the Comparative Annotation Toolkit (CAT) (Fiddes, Armstrong, et al., 2018) to annotate each of the genomes within a pangenome graph. CAT projects a reference annotation, in this case GENCODE v38, to each of the haplotypes using the underlying alignments within the graph. CAT was run on both of the Minigraph-Cactus based pangenome graphs (the GRCh38-based graph and the CHM13-based graph).

CAT was run using commit `eb2fc8752b6646f6385c12c5168dce579eb435a6`. For each graph, the autosomes were first run all together, and then the sex chromosomes were run on the

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

appropriate haplotypes. The parameters used were default parameters, except as shown below. An example CAT command run is:

```
luigi --module cat RunCat --hal=CHM13-f1g-90-mc-aug11.hal --ref-genome=GRCh38 --workers=8 --
config=cat-hprc.gencode38.autosomes.config --work-dir work-hprc-gencode38-chm13 --out-dir out-
hprc-gencode38-chm13 --local-scheduler --assembly-hub --maxCores 8 --binary-mode local >
cat.hprc.gencode38.autosomes.chm13.log
```

Analysis of 5 complex loci

Visualization of graph structures of 5 loci

We extracted subgraphs and paths of 5 loci in MC and PGGB graphs using gfabase(v0.6.0, (Lin, 2021)) and odgi (v0.6.2, (Guarracino, Heumos, et al., 2022)) respectively with the following commands:

```
gfabase sub GRCh38-f1g-90-mc-aug11.gfab GRCh38.chr1:25240000-25460000 --range --connected --view
--cutpoints 1 --guess-ranges -o RH_locus.walk.gfa
odgi extract -i chr1.pan.fa.a2fb268.4030258.6a1ecc2.smooth.og -o chr1.pan.RH_locus.og -b
chr1.RH_locus.bed -E -P
```

We then visualized the graph structures of the subgraphs using bandage (v0.8.1, (Wick et al., 2015)).

Alignment of genes to graphs

We aligned Ensembl (release 106, (Cunningham et al., 2022)) GRCh38 version gene sequences to the MC graph and PGGB graph using GraphAligner (v1.0.13, (Rautiainen & Marschall, 2020)) with parameter settings "`-x vg --try-all-seeds --multimap-score-fraction 0.1`" to identify the gene positions within the graphs.

Structural haplotypes identification

Sequences of each assembly are represented by paths in a gfa file. Structural haplotypes, such as insertions, deletions, inversions and CNVs, of each assembly are identified by tracing these paths through different big bubbles (>5,000 bp) in either MC graph or PGGB graph within those gene regions. An example command to identify big bubbles at RH locus was:

```
bcftools filter hprc-v1.0-mc-grch38.vcf.gz -r chr1:25240000-25460000 |grep LV=0 |awk '{OFS="\t";
print $1,$2,$3,$4,$5}' |tr "," "\t" |awk '{OFS="\t"; for (i=1;i<=NF;i++) {len=length($i); if
(len>5000) {print $1,$2,$3; next}}}'
```

Gene conversion detection

Since gene conversion is not shown in a form of bubbles in the graph, to reveal gene conversion events, we identified nodes that were different between a gene and its homologous gene (e.g. RHD and RHCE), which are referred to as paralogous sequence variants (PSVs). A gene conversion event is detected if a path of a gene goes through more than 4 PSVs of its homologous gene in a row

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Visualization of linear gene structures

Linear gene structures are visualized using gggenes (v0.4.1, (Wilkins, 2022)) based on structural haplotypes and gene conversions of each assembly. The length of intervals between genes is fixed (except for TMEM50A and RHCE, because those two genes are right next to each other). Lengths of genes are proportional to gene lengths in GRCh38.

Point genotyping with Giraffe/DeepVariant/DeepTrio

Alignment of reads to the pangenome

The short-reads were first split in chunks to parallelize the read mapping to the “allele-filtered graph” pangenome defined above in “Minigraph-Cactus Pangenome Pipeline”, and is identified in the dataset accompanying this paper as “clip.d9.m1000.D10M.m1000”. Mapping was performed with vg giraffe (Sirén et al., 2021) from vg release v1.37.0. For trio-based runs, the trio-sample sets of short-reads were mapped to the pangenome using vg giraffe from vg release v1.38.0. Note that the core vg algorithms for Giraffe mapping and surjection (conversion from graph space to linear space) are the same in both vg v1.37.0 and v1.38.0. The output alignments, surjected to GRCh38 in BAM format as explained below, are available at (s3://human-pangenomics/publications/PANGENOME_2022/DeepTrio/samples/) in the “bam” directory of each sample’s directory, and are organized by aligner.

Surjection to GRCh38 and indel realignment

In order to perform variant calling, GAM alignments were surjected onto the chromosomal paths from GRCh38 (chr 1-22, X, Y) using `vg surject` and the `--prune-low-cplx` option to prune short and low complexity anchors during realignment. The BAM files were sorted and split by chromosome using samtools (v1.3.1), (Li et al., 2009). The reads were realigned, first using `bamleftalign` from FreeBayes (v1.2.0), (Garrison & Marth, 2012), and then with ABRA (v2.23), (Mose et al., 2014) on target regions that were identified using *RealignerTargetCreator* from GATK (v3.8.1), (Poplin et al., 2017) and expanded by 160 nucleotides with bedtools *slop* (v2.21.0), (Quinlan & Hall, 2010).

Model training

To do variant calling with DeepVariant and DeepTrio, we trained machine learning models specific to our graph reference and vg giraffe alignment pipeline, based on our alignments. For all models, chromosome 20 was entirely held out from all input samples, to provide a control.

Training was performed on Google’s internal cluster, using unreleased Google “Tensor Processing Unit” (TPU) accelerators, from a “cold start” (i.e. without using a pre-trained model as input). We believe that nothing about the way in which we executed the training is essential to the results obtained. Cold start training is estimated to be feasible outside of the Google environment, and thus the claims we present here are falsifiable, but it is not expected to be cost-effective. Researchers looking to independently replicate our training should consider doing “warm start” training from a base model trained on other data, using commercially available

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Graphics Processing Unit (GPU) accelerators. An example procedure can be found in the DeepVariant training tutorial at (https://github.com/google/deepvariant/blob/r1.3/docs/deepvariant-training-case-study.md). We predict that this more accessible method would yield equivalent results.

For both DeepVariant and DeepTrio, the true variant calls being trained against came from the GIAB benchmark v4.2.1.

For DeepVariant, we trained on the HG002, HG004, HG005, HG006, and HG007 samples, with HG003 held out. The trained DeepVariant model is available at (s3://human-pangenomics/publications/PANGENOME_2022/DeepVariant/DEEPVARIANT_MC_Y1/).

For DeepTrio, we trained two sets of models: one on HG002, HG003, HG004, HG005, HG006 and HG007, with HG001 held out, and one on HG001, HG005, HG006, and HG007, with the HG002/3/4 trio held out. Each DeepTrio model set included parental and child models. The two trained child deeptrio models are available at (s3://human-pangenomics/publications/PANGENOME_2022/DeepTrio/models/deeptrio/child/) and (s3://human-pangenomics/publications/PANGENOME_2022/DeepTrio/models/deeptrio-no-HG002-HG003-HG004/child/), respectively. The two trained parental deeptrio models are available at (s3://human-pangenomics/publications/PANGENOME_2022/DeepTrio/models/deeptrio/parent/) and (s3://human-pangenomics/publications/PANGENOME_2022/DeepTrio/models/deeptrio-no-HG002-HG003-HG004//parent/, respectively).

Variant calling with DeepVariant

DeepVariant (v.1.3) was evaluated on HG003, using the model we trained with HG003 held out (see “Model training”). We used the `--keep_legacy_allele_counter_behavior` flag (introduced to support this analysis) and a minimum mapping quality of 1 in the `make_examples` step, before calling the variants with `call_variants`. Both VCFs and gVCFs were produced. The WDL workflow used for single sample mapping and variant calling was deposited on dockstore (*github.com/vgteam/vg_wdl/GiraffeDeepVariantLite*, 2022).

Variant calling on GRCh38 with BWA-MEM and DeepVariant

Small variants were also called using a more traditional pipeline. We aligned reads with BWA-MEM (Li, 2013) to GRCh38 with decoys but no ALTs. DeepVariant then called small variants from the aligned reads. The same version and parameters were used for DeepVariant. Only the model was changed, to the default DeepVariant model.

Variant calling with DeepTrio

Small variants were also called with DeepTrio (v1.3). For HG001, we used the DeepTrio models we trained with HG001 held out (see “Model training”). For the HG002/3/4 and HG005/6/7 trios, we used the models trained with the HG002/3/4 trio held out; the HG005/6/7 trio (except for chr20) was still included in the training set. We used the `--`

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

`keep_legacy_allele_counter_behavior` and a minimum mapping quality of 1 in the `make_examples` step before calling the variants with `call_variants`. Both VCFs and gVCFs were produced and are available at (s3://human-pangenomics/publications/PANGENOME_2022/DeepTrio/samples/) in the "vcf" directory of each sample's directory, and are organized by mapping and calling condition. The WDL workflow used for trio-based mapping and variant calling was deposited on dockstore (https://doi.org/10.5281/ZENODO.6655962).

Variant calling on GRCh38 with BWA-MEM, Dragen Graph, and DeepTrio

For DeepTrio, small variants were also called using a more traditional pipeline and a graph-based implementation of Illumina's Dragen platform (v3.7.5). The conditions evaluated were each a combination of a mapper and a reference. The *Giraffe-HPRC* condition used VG Giraffe (v1.38.0, (Sirén et al., 2021)) to align reads to the HPRC reference. The *BWA-MEM* condition used BWA-MEM (v0.7.17-r1188), (Li, 2013) to align reads to the hs38d1 human reference genome with decoys but no ALTs. The *Dragen-DeepTrioCall* condition used Illumina's Dragen platform (v3.7.5), (Miller et al., 2015) against their default graph, which was constructed using the same GRCh38 reference with decoys but no ALTs, and population contigs, SNPs and liftover sequences from datasets internal to their platform. DeepTrio then called small variants from the aligned reads. The same version and parameters were used for DeepTrio (v1.3). Only the default model was used for these conditions. We also applied the native Dragen caller and joint genotyper to the Dragen Graph based alignments for comparison purposes, referred to as *Dragen-DragenCall and Dragen-DragenJointCall,* respectively. *Dragen-DragenCall* implements a single-sample based method and is what is the default use-case for processing Dragen-graph mapped data. *Dragen-DragenJointCall* uses a pedigree-backed implementation that informs which variants are likely denovo and which are erroneous given the genotype information of the parents. In order to make a more fair comparison with Dragen, we tested these configurations to assess what implementation of Dragen variant calling can produce the best results given the available trio data.

Evaluation using the Genome in a Bottle benchmark

The small variants calls were evaluated on HG001-7 with the GIAB benchmark v4.2.1 (Wagner, Olson, Harris, Khan, et al., 2022), on HG002 in Challenging Medically-Relevant Autosomal Genes (Wagner, Olson, Harris, McDaniel, et al., 2022), and on HG002 using a preliminary draft assembly-based benchmark. For the draft assembly-based benchmark, we used dipcall to align a scaffolded, high-coverage trio-hifiasm assembly (Jarvis et al., 2022) to GRCh38 and call variants, and then we excluded structurally variant regions from the dip.bed file as described above for the benchmarking of small variants from the pangenome graph. The comparison between the callsets and truth set was made with RTG's vcfeval (Cleary et al., 2015) and Illumina's hap.py tool (Krusche et al., 2019) on confident regions of the benchmark. We used high-coverage read sets of the GIAB HG001, HG002 and HG005 trio child samples and evaluated performance within the held-out Chromosome 20 for the GIAB v4.2.1 truth set, or the whole genome for the reduced truth set of the Challenging Medically-Relevant Autosomal

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Genes. The evaluation was also stratified using the set of regions provided by the GIAB at (https://github.com/genome-in-a-bottle/genome-stratifications) (Olson et al., 2022).

Variant calls across samples from the 1000 Genomes Project

We applied our small variant calling pipeline to the high-coverage read sets for the 3,202 samples of the 1KG (1000 Genomes Project Consortium et al., 2015). The output alignments, in the GAM format, and the VCFs were saved in public buckets at (gs://brain-genomics-public/research/cohort/1KGP/vg/graph_to_grch38/). We selected 100 trios among those samples to further evaluate the quality of the calls. We tested all variants that have at least one alternate allele in a trio for Mendelian consistency. In addition, for each variant, we only considered trios where the child's genotype was different from the genotype of at least one of the parents, to minimize bias created by systematic calls (e.g. all homozygous or all heterozygous). We look at the fraction of variants-trios that fail Mendelian consistency in the whole genome and in sites that don't overlap simple repeats as defined by the "simpleRepeat" track downloaded from the UCSC Genome Browser. The results are compared with Mendelian consistency of calls provided by the 1KG that used GATK HaplotypeCaller on the reads aligned to GRCh38. We also repeated this analysis on the two trios of the GIAB v4.2.1 benchmark (HG002-7) and across the different methods of our evaluation described above (BWA-MEM, DragenGraph mappers; DeepVariant, DeepTrio, Dragen variant callers).

SV genotyping with PanGenie

VCF preprocessing

We used a VCF file created based on the snarl traversal of the MC graph as a basis for genotyping. The records contained in this VCF represent bubbles in the underlying pangenome graph as well as their nested variants, derived from the snarl tree. Each variant is marked according to their level in this tree. Variants annotated by "LV=0" correspond to the top-level bubbles. We used vcfbub (version 0.1.0, (Garrison, 2021)) with parameters `-l 0` and `-r 100000` in order to filter the VCF. This removes all non-top-level bubbles from the VCF, unless they are nested inside a top-level bubble with a reference length exceeding 100kb, i.e. top-level bubbles longer than that are replaced by their child nodes in the snarl tree. The VCF also contains the haplotypes for all 44 assembly samples, representing paths in the pangenome graph. We additionally remove all records for which more than 80% of all 88 haplotypes carry a missing allele ("."). This resulted in a set of 22,133,782 bubbles. In a next step, we used PanGenie (Ebler et al., 2022) (v1.0.0), to genotype these bubbles across all 3,202 samples from the 1KG based on high coverage Illumina reads (Byrska-Bishop et al., 2021).

Decomposition of variants

Genotyping results in genotypes for all top-level bubbles across all 1000 Genomes samples. While bi-allelic bubbles can be easily classified representing SNPs, indels or SVs, this becomes more difficult for multi-allelic bubbles contained in the VCF. Especially larger multi-allelic bubbles can contain a high number of nested variant alleles overlapping across haplotypes, represented as a single bubble in the graph. This is especially problematic when comparing the

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

genotypes computed for the whole bubble to external callsets, as coordinates of the bubble do not necessarily represent the exact coordinates of individual variant alleles carried by a sample in this region (**Supplementary Figure 25**).

In order to tackle this problem, we have implemented a decomposition approach which aims at detecting all variant alleles nested inside of multi-allelic top-level records. The idea is to detect variants from the node traversals of the reference and alternative alleles of all top level bubbles. Given the node traversals of a reference and alternative path through a bubble, our approach is to match each reference node to its leftmost occurrence in the alternative traversal, resulting in an alignment of the node traversals (**Supplementary Figure 26-a**). Nested alleles can then be determined based on insertions, deletions and mismatches in this alignment. Since the node traversals of the alternative alleles can visit the same node more than once (which is not the case for the reference alleles of the MC graph), this approach is not guaranteed to reconstruct the optimal sequence alignment underlying the nodes in these repeated regions.

As an output, the decomposition process generates two VCF files. The first one is a multi-allelic VCF which contains exactly the same variant records as the input VCF, just that annotations for all alternative alleles of a record were added to the ID tag in the INFO field. For each alternative allele, the ID tag contains IDs encoding all nested variants it is composed of, separated by a colon. The second VCF is bi-allelic and contains a separate record for each nested variant ID defining reference and alternative allele of the respective variant (**Supplementary Figure 26-b**). Both VCFs are different representations of the same genomic variation, i.e. before and after decomposition. We applied this decomposition method to the MC-based VCF file, use the multi-allelic output VCF as input for PanGenie to genotype bubbles, and use the bi-allelic VCF as well as the IDs in order to translate PanGenie's genotypes for bubbles to genotypes for all individual nested variant alleles. All downstream analyses of the genotypes are based on this bi-allelic representation (i.e. after decomposition).

While the majority of short bubbles (< 10 bp) are bi-allelic, especially large bubbles (> 1000 bp) tend to be multi-allelic. Sometimes each of the 88 haplotypes contained in the graph covers a different path through such a bubble **(Supplementary Figure 27-a)**, leading to a VCF record with 88 alternative alleles listed. We determined the number of variant alleles located inside of bi-allelic and multi-allelic bubbles in the pangenome after decomposition. As expected, the majority of SV alleles is located inside of the more complex, multi-allelic regions of the pangenome (**Supplementary Figure 27-b**).

Genotyping Evaluation based on assembly samples

We conducted a "leave-one-out" experiment in order to evaluate PanGenie's genotyping performance for the callset samples. For this purpose, we repeatedly removed one of the panel samples from the MC VCF and genotyped it using only the remaining samples as an input panel for PanGenie. We later used the genotypes of the left-out sample as ground truth for evaluation. We repeated this experiment for five of the callset samples (HG00438, HG00733, HG02717, NA20129 and HG03453) using 1000 Genomes high coverage Illumina reads (Byrska-Bishop et

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

al., 2021). PanGenie is a re-genotyping method. Therefore, like every other re-typer, it can only genotype variants contained in the input panel VCF, that is, it is not able to detect variants unique to the genotyped sample. For this reason we removed all variant alleles (after decomposition) unique to the left-out sample contained in the truth set for evaluation. In order to evaluate the genotype performance, we used the weighted genotype concordance (Ebler et al., 2022). (**Supplementary Figure 28)** shows the results stratified by different regions. (**Supplementary Figure 28-A)** shows concordances in biallelic and multiallelic regions of the MC VCF. The biallelic regions include only bubbles with two branches. The multiallelic regions include all bubbles in which haplotypes cover more than two different paths. (**Supplementary Figure 28-B)** shows the same results stratified by genomic regions defined by GIAB that we obtained from:
easy: (https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/genome-stratifications/v3.0/GRCh38/union/GRCh38_notinalldifficultregions.bed.gz)
low-mappability: (https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/genome-stratifications/v3.0/GRCh38/union/GRCh38_alllowmapandsegdupregions.bed.gz)
repeats: (https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/genome-stratifications/v3.0/GRCh38/LowComplexity/GRCh38_AllTandemRepeats_gt100bp_slop5.bed.gz)
other-difficult: (https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/genome-stratifications/v3.0/GRCh38/OtherDifficult/GRCh38_allOtherDifficultregions.bed.gz)

Here and in the following, we consider results for SNPs, indels (1-49bp), SV deletions, SV insertions and other SV alleles, defined as follows: SV deletions include all alleles for which length(REF) >= 50bp and length(ALT) = 1. SV insertions include all alleles for which length(REF) = 1 and length(ALT) >= 50. All other alleles with a length >= 50bp are included in “others”.

Overall, weighted genotype concordances are high for all variant types. Especially variant alleles in biallelic regions of the graph are very well genotypable. Alleles inside of multiallelic bubbles are more difficult to genotype correctly since PanGenie needs to decide between several possible alternative paths, while there is only two such paths for biallelic regions **(Supplementary Figure 28-A)**. Furthermore, genotyping accuracy depends on the genomic context (**Supplementary Figure 28-B**). Regions with low mappability, repetitive regions and other difficult regions are harder to genotype than regions classified as “easy” by GIAB.

Creating a high quality subset

We generated genotypes for all 3,202 1000 Genomes samples with PanGenie and defined a high quality subset of SV alleles that we can reliably genotype. For this purpose, we applied a machine learning approach similar to what we have presented previously (Ebert et al., 2021; Ebler et al., 2022). We define positive and negative subsets of variants based on the following filters:

- **ac0_fail:** a variant allele was genotyped with an AF of 0.0 across all samples

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

- **mendel_fail:** the mendelian consistency across trios is less than 80% for a variant allele. Here, we use a strict definition of mendelian consistency which excludes all trios with only 0/0, only 0/1 and only 1/1 genotypes.
- **gq_fail:** less than 50 high quality genotypes were reported for this variant allele
- **self_fail:** genotyping accuracy of a variant allele across the panel samples is less than 90%
- **nonref_fail:** not a single non-0/0 genotype was genotyped correctly across all panel samples

The positive set includes all variant alleles that passed all five filters. The negative set contains all variant alleles that passed the *ac0_fail* filter but failed at least three of the other filters. We trained a support vector regression (SVR) approach based on these two sets that uses multiple features including allele frequencies, mendelian consistencies or the number of alternative alleles transmitted from parents to children. We applied this method to all remaining variant alleles genotyped with an AF > 0, resulting in a score between -1 (bad) and 1 (good) for each. We finally define a filtered set of variants which includes the positive set, as well as all variant alleles with a score of >= -0.5.

We show the number of variant alleles contained in the unfiltered set, the positive set as well as the filtered set in **Supplementary Table 16**. Since our focus is on SVs and since 65% of all SNPs and indels are already contained in the positive set, we applied our machine learning approach only to SVs. We found that 50%, 33% and 26% of all deletion, insertion and “other” alleles were contained in the final, filtered set of variants, respectively. Note that these numbers take all distinct SV alleles contained in the callsets into account. Especially for insertions and “other” SVs, many of these alleles are very similar, with sometimes only a single base pair differing. Therefore, it is likely that many of these actually represent the same events. Our genotyping and filtering approach helps to remove such redundant alleles.

In order to evaluate the quality of the PanGenie genotypes, we compared the allele frequencies observed for the SV alleles across all 2,504 unrelated 1000 Genomes samples to their allele frequencies observed across the 44 assembly samples in the MC callset. (**Supplementary Figures 29, 30, and 31)** show the results for SV deletions, insertions and other SV alleles. We observed that the allele frequencies between both sets match well, resulting in correlations (pearson) of 0.93, 0.87 and 0.81 for deletions, insertions and “other” alleles contained in the unfiltered set. For the filtered set, we observed correlations of 0.96, 0.93 and 0.90, respectively. We also analyzed the heterozygosity of the PanGenie genotypes across all 2,504 unrelated 1000 Genomes samples and observed a relationship close to what is expected by Hardy-Weinberg equilibrium (**Supplementary Figures 29, 30, and 31**, lower panel).

Number of SVs per sample

We compared our filtered set of variant alleles to the HGSVC PanGenie genotypes (v2.0 “lenient” set, (Ebert et al., 2021)) and Illumina-based SV genotypes (Byrska-Bishop et al., 2021). A direct comparison of the three callsets is difficult. The HGSVC and HPRC callsets are based on variant calls produced from haplotype-resolved assemblies of 32 and 44 samples,

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

respectively (Ebert et al., 2021). For each callset, variants were re-genotyped across all 3,202 1000 Genomes samples. Note that the callset samples for HPRC and HGSVC are disjoint. Since re-genotyping cannot discover novel variants, both callsets will miss variants carried by 3,202 samples that were not seen in the assembly samples. In contrast, the 1KG callset contains short-read based variant calls produced for each of the 3,202 1000 Genomes samples. Another difference between the HGSVC and HPRC callsets is that in the HGSVC callset, highly similar alleles are merged into a single record to correct for representation differences across different samples or haplotypes. The HPRC callset however, keeps all these alleles separately even if there is only a single basepair difference between them. To make the callsets better comparable, we merged clusters of highly similar alleles in the HPRC filtered set prior to comparisons with other callsets. This was done with truvari ((English et al., 2022), version v3.1.0) using the command: `truvari collapse -r 500 -p 0.95 -P 0.95 -s 50 -S 100000`.

In order to be able to properly compare the callsets despite their differences, we counted the number of SV alleles present in each sample (genotype 0/1 or 1/1) in each callset and plotted the corresponding distributions stratified by genome annotations from GIAB (same as above, **Figure 6D**). We also generated the same plot including only common SV alleles with an AF > 5% across all 3,202 samples (**Supplementary Figure 34**). Both plots show that both assembly-based callsets (HPRC, HGSVC) are able to access more SVs across the genome than the short-read-based 1KG callset, especially deletions < 300bp and insertions (**Figure 6E**). This confirms that SV callers based on short-reads alone miss a large portion of SVs located in regions inaccessible by short-read alignments, which has been reported previously by several studies (Ebert et al., 2021; Zhao et al., 2021). In the “easy” regions, the number of SVs per sample is consistent across all three callsets. For the other regions however, results indicate that the HPRC filtered genotypes give access to more variant alleles than the HGSVC lenient set, especially insertions and variants in regions of low mappability and tandem repeats (**Figure 6D, Figure 6E**).

Evaluation based on medically relevant SVs

In addition to all 3,202 1000 Genomes samples we also genotyped sample HG002 based on Illumina reads from (J. M. Zook et al., 2016). We used the GIAB CMRG benchmark containing medically relevant SVs (Wagner, Olson, Harris, McDaniel, et al., 2022), downloaded from: (https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/AshkenazimTrio/HG002_NA24385_son/CMRG_v1.00/GRCh38/StructuralVariant/) for evaluation. Like for the 1000 Genomes samples, we used the MC-based VCF (see above) containing variant bubbles and haplotypes of 44 assembly samples as an input panel for PanGenie. We extracted all variant alleles with a length >= 50bp from our genotyped VCF (biallelic version, after decomposition). We converted the ground truth VCF into a biallelic representation using `bcftools norm -m -any` and kept all alleles with length >= 50bp. We used truvari ((English et al., 2022), version v3.1.0) with parameters `--multimatch --includebed <medically-relevant-sv-bed> -r 2000 --no-ref a -C 2000 --passonly` in order to compare our genotype predictions to the medically relevant SVs. Results are shown in **(Supplementary Table 18, left)**. Since PanGenie

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

is a re-typing method, it can only genotype variants provided in the input and thus cannot detect novel alleles. Since HG002 is not among the panel samples, the input VCF misses variants unique to NA24385. Thus, these unique variants cannot be genotyped by PanGenie and will be counted as false negatives during evaluation. Therefore, we computed an “adjusted” version of the recall which excludes SV alleles unique to HG002 (i.e. alleles not in the graph) from the truth set for evaluation. In order to identify which SV alleles were unique, we compared each of the 44 panel samples to the ground truth VCF using truvari in order to identify the false negatives for each sample. Then we computed the intersection of false negative calls across all samples. The resulting set then contains all variant alleles unique to the HG002 ground truth set. We found 15 such unique SV alleles among the GIAB CMRG variants. We removed these alleles from the ground truth set and recomputed precision/recall statistics for our genotypes. Adjusted precision/recall values are shown in **(Supplementary Table 18, right).**

Read mapping at VNTR regions

Simulating and mapping VNTR reads

Raw VNTR coordinates on GRCh38 (chr1-22 and sex chromosomes only) were generated using TRF v4.09 (Benson, 1999) with command: `trf hg38.fa 2 7 7 80 10 50 500 -f -d`. Only repeats with period size between (6,10k) bp, total length > 100bp and not overlapping with centromeric regions were selected, leaving a total number of 98,021 non-overlapping loci. Using the raw VNTR coordinates on GRCh38 as input, VNTR regions across 96 haplotypes (including GRCh38) were annotated using the build module in danbing-tk v1.3 (Lu et al., 2021) (dist_scan=700, dist_merge=1, TRwindow=100000, MBE_th1=0.3, MBE_th2=0.6).

Whole-genome paired-end error-free short-reads were simulated at ~30x for each genome, or equivalently ~15x for each haplotype. A read pair was generated for every 20 bp with fragment size = 500 bp and read length = 150bp. Paired-end read mapping to the mc graph was done using vg giraffe v1.39.0 (Sirén et al., 2021) using the command `vg giraffe -x $pref.xg -g $pref.gg -H $pref.gbwt -m $pref.min -d $pref.dist -p -f <(zcat $h1 $h2) -i -t 16`) while mapping to GRCh38 was done using bwa-mem v0.7.17-r1188 (Li, 2013) using the command `bwa mem -t 16 -Y -K 100000000 -p $ref <(zcat $h1 $h2)`. For a fair comparison, GRCh38 plus decoy minus alt/HLA contigs were used as reference to match the paths included in the mc graph.

Evaluating read mapping accuracy at VNTR regions

To evaluate the performance of read mapping using the mc graph plus giraffe, the VNTR information from danbing-tk were used to annotate each node in the graph by traversing each haplotype path. Every node that covers a VNTR region has a tuple that denotes the intersected interval; any aligned reads overlapping with the interval were considered mapped to the VNTR. Similarly, a read simulated from an interval overlapping with a VNTR was considered derived from the VNTR. To evaluate the performance of GRCh38 plus bwa-mem, the mapped region by each read was obtained using the bamtobed submodule in bedtools v2.30.0 (Quinlan & Hall, 2010). The VNTR annotations on GRCh38 were used to determine whether a read was mapped

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

to a VNTR.

For each read, we tracked its source and mapped VNTR(s), and used this information to compute accuracy . Only VNTRs present in danbing-tk's annotations were tracked; otherwise they were labeled "untracked" the same as non-VNTR regions. A true positive denotes mapping from a VNTR to its original VNTR. An exogenous false positive denotes mapping from untracked regions to a VNTR. An endogenous false positive denotes mapping from a VNTR to another VNTR. A false negative denotes mapping from untracked regions to untracked regions. Any alignments in the json output of giraffe that did not contain the "mapping" field were considered unmapped. The two ends of a read pair that did not map to the same chromosome by bwa-mem were also considered unmapped.

Estimating VNTR length variants from read depths

The whole-genome sequencing (WGS) samples for 35 genomes (HG00438, HG00621, HG00673, HG00733, HG00735, HG01071, HG01106, HG01109, HG01175, HG01243, HG01258, HG01361, HG01891, HG01928, HG01952, HG01978, HG02055, HG02080, HG02145, HG02148, HG02257, HG02572, HG02622, HG02630, HG02717, HG02723, HG02818, HG02886, HG03098, HG03453, HG03486, HG03492, HG03579, NA18906, NA19240) were mapped to the mc graph using vg giraffe as described in the "Point genotyping with Giraffe/DeepVariant/DeepTrio" section. Using the VNTR annotations described in the previous section, the number of reads mapped to each VNTR region in the mc graph was calculated as a proxy for VNTR length. VNTRs with invariant length across the 35 genomes were removed from analysis, leaving a total of 60,861 loci.

As a baseline control, the read depth of each VNTR region for the 35 WGS samples produced by mapping reads to GRCh38 was also computed with mosdepth v0.3.1 (Pedersen & Quinlan, 2018) using the command `mosdepth -t 4 -b $VNTR_bed -x -f $hg38 $pref $cram`. To be able to compare with the graph-based approach, VNTRs with missing annotation on GRCh38 were further removed, leaving a total of 60,386 VNTRs.

RNA-seq mapping evaluation

We augmented the "allele-filtered graph" (see Point genotyping section) with edges for splice junctions to create a spliced pangenome graph using the rna subcommand in the vg toolkit v1.38.0 with a maximum node length set to 32 (`vg rna -k 32`) (Sibbesen et al., 2021). The transcript annotations that were used to define the splice junctions consisted of the CAT transcript annotations on each assembly together with splice junctions from the GENCODE (v38) annotation (Frankish et al., 2021). Transcripts from the GENCODE (v38) annotation were further added as paths to the spliced pangenome graph. For comparison, we created two other references/graphs. A spliced reference constructed from the reference sequence, and a spliced pangenome graph constructed from the high coverage 1KG (1000GP) phased variant set created by the New York Genome Center (https://www.internationalgenome.org/data-portal/data-collection/30x-grch38) (Byrska-Bishop et al., 2021). Both were constructed using the GENCODE (v38) transcript annotation, and a minimum AF filter of 0.001 was used for the

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

1000GP graph. For chromosomes Y and MT, we used the unphased 1000GP variant set and only included variants that passed all filters. In addition, the trio NA12878, NA12891 and NA12892 were filtered from the 1000GP variant set since RNA-seq data from NA12878 were used for the mapping evaluation. BCFtools v1.9 was used for filtering the variant sets (Danecek et al., 2021). For each graph we created the indexes needed for mapping using the vg toolkit v1.38.0 with default parameters, except when pruning where edges on embedded paths were restored (`vg prune -r`). Furthermore, for the spliced HPRC pangenome graph it was necessary to use stricter pruning parameters (`vg prune -r -k 64 -M 64`). For the spliced reference we created the index needed by the RNA-seq mapper STAR using default parameters.

We simulated RNA-seq reads with a pipeline that was designed to preserve complex genome variation in the simulated data. The transcript sequences used for the simulation were derived from the GENCODE (v38) transcript annotations projected onto assembled haplotypes from HG002. Specifically, we used minigraph-cactus to create an alignment between GRCh38 and the two HG002 haplotypes, which were held out of the main pangenome graph for benchmarking. We then used CAT to lift the transcript annotations over to these haplotypes. We constructed a spliced personal genome graph using the `vg rna` subcommand, and then we simulated reads using `vg sim` (commit 2cea1e2), using an Illumina NovaSeq cDNA read set (SRR18109271) to fit model parameters. This essentially amounts to simulating directly from the projected transcript sequences. The transcripts were simulated with uniform expression, split evenly between the two haplotypes. This expression profile is not biologically realistic, but it avoids the difficulty of choosing a particular expression profile as representative for all tissues and stages of development. Moreover, existing estimated profiles would be biased toward the tools that were used to estimate them. We simulated 5,000,000 paired-end 150bp RNA-seq reads.

Both simulated and real Illumina RNA-seq reads were mapped to the graphs using `vg mpmap` (commit 2cea1e2) with default parameters. In addition, the reads were mapped to the spliced reference using STAR (version 2.7.10a) with default parameters (Dobin et al., 2013). For the real data we used NA12878 RNA-seq data from both Tilgner, et al. (SRR1153470) (Tilgner et al., 2014) and the ENCODE project (ENCSR000AED, replicate 1) (Davis et al., 2018; ENCODE Project Consortium, 2012).

We used the same approach as described in Sibbesen, et al. (Sibbesen et al., 2021) to evaluate the alignments. Briefly, for the simulated data the graph alignments were compared to the truth alignments by estimating their overlap on the reference genome paths. The graph alignments were projected to the reference paths using `vg surject -S`. An alignment was considered correct if it overlapped 90% of the truth alignment. For the real data, the average read coverage of each exon on the reference path calculated from the projected graph alignments were compared to the corresponding coverages estimated from long-read alignments. For the long-read data we used PacBio Iso-Seq alignments from the ENCODE project (ENCSR706ANY, all replicates), which come from the same cell line as the Illumina data. The long-read alignments were used to define the exons, and only primary long-read alignments with a mapping quality of

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

at least 30 were used. The alignments for the four Iso-Seq replicates (ENCFF247TLH, ENCFF431IOE, ENCFF520MMC, ENCFF626GWM) were combined and filtered using SAMtools v1.15 (Danecek et al., 2021). BEDTools v2.30.0 was used to convert the alignments to exons coordinates (Quinlan & Hall, 2010).

The scripts that was used for graph construction, read simulation, mapping and evaluation are available at (https://github.com/jonassibbesen/hprc-rnaseq-analyses-scripts).

ChIP-seq analysis

We aligned H3K4me1, H3K27ac and ATAC-seq obtained from monocyte-derived macrophages from 30 individuals (Groza et al., 2022) using *vg map* (Garrison et al., 2018) to the hg38 reference genome graph and to the HPRC genome graph. Then, we called peaks using *Graph Peak Caller v1.2.3 (Grytten et al., 2019)* on both sets of alignments for each of the 30 H3K4me1, H3K27ac, and ATAC-seq samples. To identify HPRC-only peaks, we projected HPRC coordinates to the hg38 path using *Graph Peak Caller* and compared intervals using *bedtools (Quinlan & Hall, 2010)*. We named HPRC peaks that overlap hg38 peaks as common peaks and those that do not as HPRC-only. We calculated the expected frequency (as inverse cumulative distributions) of common and HPRC-only peaks among the 30 samples by resampling the peaks of each sample from the peaks of all the samples and re-counted the number of overlaps. We repeated this simulation 100 times and plotted the average curves. We determined heterozygous variants in our samples by aligning WGS datasets for each sample to the HPRC graph using *vg map* and genotyping the variants using *vg call -a*. We narrowed the list of heterozygous SVs above 50 bp in each sample with the aim of looking for allelic-specific peaks. For each epigenomic sample, we obtained allelic-specific read counts within peaks that lie on the previously identified loci by running *vg call -a* on the epigenomic HPRC alignments, which outputs the numbers of reads on each path in a bubble (DP and AD fields in the VCF output). We then assigned peaks to the SV or reference allele, or both alleles with a two-tailed binomial test parameterized on the sum of reads on both alleles and $p = 0.05$. Any peak with reads on one allele, but not on the other was assigned to the allele with the reads. Read counts were proportionally adjusted for the difference in length between the reference and SV alleles. Processed data, scripts and code for the above steps are available on Zenodo (Groza & Bourque, 2022).

Population genetic analyses

Overview

Although our pangenome contains a relatively small number of genomes (N=44), the high quality and contiguity of its component assemblies in principle allows us to consider variation in regions that were previously inaccessible to study. We thus undertook a basic survey of population genetic features relative to chromosomal regions. We used the PGGB graph due to its inclusion of all input contigs and our ability to apply Flagger assembly annotations to its embedded paths. We generated a T2T-CHM13 based VCF from the subset of the PGGB graph containing only Flagger confident regions. For each chromosome-scale region, both whole

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

chromosomes and the short (p) and long (q) arms, we computed a PCA and evaluated the first two principal components (**Supplementary Figure 50**). Labeling genomes by 1000 Genomes population revealed a consistent set of superpopulation clusters in metacentric p- and q-arms, and in acrocentrics q-arms. However, in acrocentric p-arms, traditional population stratification breaks down. A quantitative evaluation based on k-means clustering (**Supplementary Figure 51**) shows that the optimal number of clusters is significantly lower (Wilcoxon p = 0.013) on the acrocentric than metacentric p-arms, a distinction not found elsewhere (**Supplementary Figure 52**). Although our restriction of the analysis to Flagger confident regions should mitigate effects of assembly error, this pattern may represent alignment error or difficulty in chromosome assignment related to patterns of “misjoins” observed in the initial assemblies (**Figure 1B**). However, an underlying biological mechanism driving homogenization between these regions cannot be ruled out, and would be consistent with prior studies of satellite sequences in these regions (Choo et al., 1988, 1989). In total, this indicates significant difficulty in utilizing these regions for population genetic study.

Analysis approach

Although the size of the population sample represented in our pangenome is small, it provides unprecedented access to previously under-ascertained regions of the genome. We sought to understand the potential utility of these regions for future population genetic studies using regional principal components analysis (PCA) based on variants called vs. the CHM13 and GRCh38 references. For these analyses we considered both the PGGB (whole pangenome, combined) and minigraph/cactus (reference-based, distinct CHM13 and GRCh38) graphs. For both graph models, the CHM13 VCFs provide access to regions that were not previously observed by studies based on GRCh38, where short-read based studies may have difficulty reliably aligning and calling variants. In combination, these two graphs provide cross-validation of implied population genetic patterns in these new regions, which we explore here.

To understand chromosome-specific patterns of variation, we applied PCA to each autosomal chromosome independently, to the VCFs from PGGB (PGGB-CHM13, PGGB-GRCh38).
To ensure that observed patterns were not derived from higher rates of assembly error in the repetitive regions of acrocentric p-arms, we used our Flagger confident region annotations to prune the PGGB graph (using odgi inject to inject the confident regions as subpaths and then odgi prune to remove the full original paths that were including unreliable regions) to only confident regions of assemblies. We then reapplied vg deconstruct to this graph to obtain a new set of SNPs (the code for the PGGB graphs pruning and variant calling on the pruned graphs can be found at the following link:
https://github.com/pangenome/HPRCyear1v2genbank/blob/main/workflows/confident_variants.md). Genome-wide, we find that pruning reduced the number of called SNPs by only 1.188% (previous N=23,272,652, pruned N=22,996,113). The total reduction in the acrocentrics was higher, with 6.29% fewer SNPs (previous N=3,735,605, pruned N=3,676,746), indicating difficulty in assembling these regions. We note that the PCA sample distributions remain virtually identical (not shown), indicating that the patterns observed in the full graph are maintained despite assembly issues. In these filtered PGGB-CHM13 and PGGB-GRCh38 VCFs, We considered all biallelic SNPs relative to the chosen reference, regardless of variant

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

nesting level (not shown: filtering for only SNPs LV=0 or LV>0 yielded nearly identical results). A qualitative evaluation suggested no significant differences in PCA patterns across the metacentric chromosomes (**Supplementary Figure 50**). However, in the p-arms of the acrocentrics (chr13, chr14, chr15, chr21, and chr22), which are accessible in the PGGB-CHM13 VCF, we observed a reduction in population differentiation and a higher rate of variance explained in the lowest principal component.

To investigate this quantitatively, we measured the number of clusters implied by the PCA for the PGGB-CHM13 VCFs, using K-Means Clustering to automatically determine the optimal number of clusters for each PCA ("gap_stat" clustering in the fviz_nbclust function of the factoextra R package) (analysis code: https://github.com/SilviaBuonaiuto/hprcPopGenAnalysis). Applying this approach to 3 PCAs per chromosome VCF, we obtain optimal cluster counts for the p-arm, q-arm, and whole chromosome. In metacentric chromosomes, we usually observe optimal numbers of clusters approximately corresponding to the number of expected world population groupings in the input genomes (3-5, as in **Supplementary Figure 51**). However, in the p-arms of the acrocentrics, we observe many fewer, in general only one cluster, indicative of reduced population differentiation compared to other parts of the acrocentric chromosomes. This pattern is only apparent in the PGGB graph based on CHM13. To evaluate the difference quantitatively, we apply a Wilcoxon rank-sum test to compare the differences between cluster count distributions in metacentric vs. acrocentric chromosomes across the whole chromosome, the q-arm, and the p-arm. We find insignificant differences between the distributions between acrocentric and metacentric chromosomes at a chromosome scale, and in the q-arms, but a significant difference (Wilcoxon $p = 0.013$) in the case of acrocentric p-arms (**Supplementary Figure 52**).

This analysis indicates that significant challenges remain for the use of these new regions in population genetic studies. Patterns observed in PCA projections of the pangenome across all chromosomes suggests a distinct process of variation sharing between populations within the short arms of the acrocentrics. In effect, we observe a more homogenous population in these regions when using the CHM13 assembly as a reference. This reference contains real sequences in these regions while GRCh38 contains gaps which render analysis impossible. The apparent population homogenization could be driven by error. We have mitigated this issue by utilizing only SNPs found in Flagger confident regions, but this does not guard against potential sources of alignment error that are likely to be amplified by the repetitive sequences in these loci. It is additionally possible that the chromosome specific partitioning process applied by both graph models is failing to correctly partition contigs on these short arms. The known homology between the short arms bolsters the possibility of ongoing sequence information exchange between non-homologous chromosomes (Nurk et al., 2022), which would be consistent with the patterns we observe. In sum, this analysis shows that, when using CHM13 as a reference, the behavior of sequences on the short arms of the acrocentrics in the PGGB graph is not similar to that of other sequences in the pangenome.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Human Pangenome Reference Consortium Authors

Wen-Wei Liao[1,2,3,*], Mobin Asri[4,*], Jana Ebler[5,*], Daniel Doerr[5], Marina Haukness[4], Glenn Hickey[4], Shuangjia Lu[3], Julian K. Lucas[4], Jean Monlong[4], Haley J. Abel[6], Silvia Buonaiuto[7], Xian H. Chang[4], Haoyu Cheng[8,9], Justin Chu[8], Vincenza Colonna[7,10], Jordan M. Eizenga[4], Xiaowen Feng[8,9], Christian Fischer[10], Robert S. Fulton[1], Shilpa Garg[11], Cristian Groza[12], Andrea Guarracino[13], William T Harvey[14], Simon Heumos[15,16], Kerstin Howe[17], Miten Jain[18], Tsung-Yu Lu[19], Charles Markello[4], Fergal J. Martin[20], Matthew W. Mitchell[21], Katherine M. Munson[14], Moses Njagi Mwaniki[22], Adam M. Novak[4], Hugh E. Olsen[4], Trevor Pesout[4], David Porubsky[14], Pjotr Prins[10], Jonas A. Sibbesen[23], Chad Tomlinson[1], Flavia Villani[10], Mitchell R. Vollger[14,24], Lucinda L Antonacci-Fulton[1], Gunjan Baid[34], Carl A. Baker[14], Anastasiya Belyaeva[34], Konstantinos Billis[20], Andrew Carroll[34], Pi-Chuan Chang[34], Sarah Cody[1], Daniel E. Cook[34], Omar E. Cornejo[35], Mark Diekhans[4], Peter Ebert[5], Susan Fairley[20], Olivier Fedrigo[36], Adam L. Felsenfeld[37], Giulio Formenti[36], Adam Frankish[20], Yan Gao[38], Carlos Garcia Giron[20], Richard E. Green[39,40], Leanne Haggerty[20], Kendra Hoekzema[14], Thibaut Hourlier[20], Hanlee P. Ji[41], Alexey Kolesnikov[34], Jan O. Korbel[42], Jennifer Kordosky[14], HoJoon Lee[41], Alexandra P. Lewis[14], Hugo Magalhães[5], Santiago Marco-Sola[43,44], Pierre Marijon[5], Jennifer McDaniel[29], Jacquelyn Mountcastle[36], Maria Nattestad[34], Nathan D. Olson[29], Daniela Puiu[45], Allison A Regier[1], Arang Rhie[28], Samuel Sacco[46], Ashley D. Sanders[47], Valerie A. Schneider[48], Baergen I. Schultz[37], Kishwar Shafin[34], Jouni Sirén[4], Michael W. Smith[37], Heidi J. Sofia[37], Ahmad N. Abou Tayoun[49,50], Françoise Thibaud-Nissen[48], Francesca Floriana Tricomi[20], Justin Wagner[29], Jonathan M. D. Wood[17], Aleksey V. Zimin[45,51], Alice B. Popejoy[52], Guillaume Bourque[25,26,27], Mark JP Chaisson[19], Paul Flicek[20], Adam M. Phillippy[28], Justin M. Zook[29], Evan E. Eichler[14,30], David Haussler[4,30], Erich D. Jarvis[31,30], Karen H. Miga[4], Ting Wang[32], Erik Garrison[10,+], Tobias Marschall[5,+], Ira Hall[3,33,+], Heng Li[8,9,+], Benedict Paten[4,+]

* These authors contributed equally
+ Corresponding authors: egarris5@uthsc.edu, tobias.marschall@hhu.de, ira.hall@yale.edu, hli@jimmy.harvard.edu, bpaten@ucsc.edu

1 McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO 63108, USA
2 Department of Medicine, Washington University School of Medicine, St. Louis, MO 63110, USA
3 Department of Genetics, Yale University School of Medicine, New Haven, CT 06510, USA
4 UC Santa Cruz Genomics Institute, University of California, Santa Cruz, 1156 High St, Santa Cruz, CA, USA
5 Institute for Medical Biometry and Bioinformatics, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany
6 Division of Oncology, Department of Internal Medicine, Washington University School of Medicine, St. Louis, MO 63110, USA
7 Institute of Genetics and Biophysics, National Research Council, Naples 80111, Italy
8 Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA 02215, USA
9 Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02215, USA

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

10 Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, TN 38163, USA
11 Department of Biology, University of Copenhagen, Denmark
12 Quantitative Life Sciences, McGill University, Montreal, Québec H3A 0C7, Canada
13 Genomics Research Centre, Human Technopole, Milan 20157, Italy
14 Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA 98195, USA
15 Quantitative Biology Center (QBiC), University of Tübingen, Tübingen 72076, Germany
16 Biomedical Data Science, Department of Computer Science, University of Tübingen, Tübingen 72076, Germany
17 Tree of Life, Wellcome Sanger Institute, Hinxton, Cambridge, CB10 1SA, UK
18 Northeastern University, Boston, MA 02115, USA
19 University of Southern California, Quantitative and Computational Biology, Los Angeles, CA, USA
20 European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Cambridge, CB10 1SD, UK
21 Coriell Institute for Medical Research, Camden, NJ 08103, USA
22 Department of Computer Science, University of Pisa, Pisa 56127, Italy
23 Center for Health Data Science, University of Copenhagen, Denmark
24 Division of Medical Genetics, University of Washington School of Medicine, Seattle, WA 98195, USA
25 Department of Human Genetics, McGill University, Montreal, Québec H3A 0C7, Canada
26 Canadian Center for Computational Genomics, McGill University, Montreal, Québec H3A 0G1, Canada
27 Institute for the Advanced Study of Human Biology (WPI-ASHBi), Kyoto University, Kyoto 606-8501, Japan
28 Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA
29 Material Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, MD 20877, USA
30 Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA
31 The Rockefeller University, New York, NY 10065, USA
32 Department of Genetics, Washington University School of Medicine, St. Louis, MO 63110, USA
33 Center for Genomic Health, Yale University School of Medicine, New Haven, CT 06510, USA
34 Google LLC, 1600 Amphitheater Pkwy, Mountain View, CA 94043, USA
35 School of Biological Sciences, Washington State University, Pullman WA 99163, USA
36 The Vertebrate Genome Laboratory, The Rockefeller University, New York, NY 10065, USA
37 National Institutes of Health (NIH)–National Human Genome Research Institute, Bethesda, MD, USA
38 Center for Computational and Genomic Medicine, The Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA.
39 Department of Biomolecular Engineering, University of California, Santa Cruz, 1156 High St., Santa Cruz, CA 95064, USA

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

40 Dovetail Genomics, Scotts Valley, CA 95066, USA
41 Division of Oncology, Department of Medicine, Stanford University School of Medicine, Stanford, CA, 94305, USA
42 European Molecular Biology Laboratory, Genome Biology Unit, Meyerhofstr. 1, 69117 Heidelberg, Germany
43 Computer Sciences Department, Barcelona Supercomputing Center, Barcelona, Spain
44 Departament d'Arquitectura de Computadors i Sistemes Operatius, Universitat Autònoma de Barcelona, Barcelona, Spain
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 Al Jalila Genomics Center of Excellence, Al Jalila Children's Specialty Hospital, Dubai, UAE
50 Center for Genomic Discovery, Mohammed Bin Rashid University of Medicine and Health Sciences, Dubai, UAE
51 Center for Computational Biology, Johns Hopkins University, Baltimore, MD 21218, USA
52 Department of Public Health Sciences, University of California, Davis, One Shields Avenue, Medical Sciences 1C, Davis, CA 95616

Author Contributions

Pangenome empirical analysis / pangenome QC

Wen-Wei Liao, Daniel Doerr, Marina Haukness, Glenn Hickey, Jean Monlong, Haley J. Abel, Justin M. Zook, Evan E. Eichler, Tobias Marschall, Ira Hall, Pierre Marijon, Justin Wagner

Paper writing

Wen-Wei Liao, Mobin Asri, Jana Ebler, Daniel Doerr, Marina Haukness, Glenn Hickey, Shuangjia Lu, Jean Monlong, Shilpa Garg, Erik Garrison, Tsung-Yu Lu, Matthew W. Mitchell, Adam M. Novak, Trevor Pesout, Jonas A. Sibbesen, Mitchell R. Vollger, Guillaume Bourque, Karen H. Miga, Tobias Marschall, Ira Hall, Benedict Paten, Robert S. Fulton, Richard E. Green, Leanne Haggerty, Hugh E. Olsen, Fergal J. Martin

Paper editing

Wen-Wei Liao, Daniel Doerr, Marina Haukness, Glenn Hickey, Xian H. Chang, Haoyu Cheng, Andrea Guarracino, Erik Garrison, Adam M. Novak, Pjotr Prins, Adam M. Phillippy, Evan E. Eichler, Erich D. Jarvis, Karen H. Miga, Tobias Marschall, Ira Hall, Heng Li, Benedict Paten, Omar E. Cornejo, Peter Ebert, Giulio Formenti, Ahmad N. Abou Tayoun, Aleksey V. Zimin

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Assembly creation

Mobin Asri, Julian K. Lucas, Haoyu Cheng, Adam M. Phillippy, Heng Li, Daniela Puiu, Allison A Regier, Aleksey V. Zimin

Assembly QC / Assembly reliability analysis

Mobin Asri, Julian K. Lucas, Haoyu Cheng, Justin Chu, Shilpa Garg, Trevor Pesout, David Porubsky, Chad Tomlinson, Mitchell R. Vollger, Adam M. Phillippy, Justin M. Zook, Evan E. Eichler, Karen H. Miga, Heng Li, Richard E. Green, Kerstin Howe, Jennifer McDaniel, Nathan D. Olson, Daniela Puiu, Allison A Regier, Arang Rhie, Valerie A. Schneider, Kishwar Shafin, Françoise Thibaud-Nissen, Justin Wagner, Jonathan M. D. Wood

Pangenome applications: structural variants

Jana Ebler, Glenn Hickey, Haley J. Abel, William T Harvey, Pjotr Prins, Erik Garrison, Evan E. Eichler, Tobias Marschall, Hanlee P. Ji, Hugo Magalhães

Pangenome graph creation

Daniel Doerr, Glenn Hickey, Andrea Guarracino, Simon Heumos, Moses Njagi Mwaniki, Flavia Villani, Yan Gao, Santiago Marco-Sola, Erik Garrison

Data coordination / management

Marina Haukness, Julian K. Lucas, William T Harvey, Chad Tomlinson, Adam M. Phillippy, Erich D. Jarvis, Karen H. Miga, Ting Wang, Lucinda L Antonacci-Fulton, Sarah Cody, Mark Diekhans, Susan Fairley, Robert S. Fulton, Richard E. Green, Miten Jain

Transcriptome / annotation

Marina Haukness, Jordan M. Eizenga, Fergal J. Martin, Mitchell R. Vollger, Mark JP Chaisson, Konstantinos Billis, Mark Diekhans, Adam Frankish, Carlos Garcia Giron, Leanne Haggerty, Thibaut Hourlier, Francesca Floriana Tricomi

Pangenome applications: small variants

Glenn Hickey, Jean Monlong, Adam M. Novak, Pjotr Prins, Justin M. Zook, Gunjan Baid, Anastasiya Belyaeva, Andrew Carroll, Pi-Chuan Chang, Daniel E. Cook, Hanlee P. Ji, Alexey Kolesnikov, Maria Nattestad, Kishwar Shafin, Justin Wagner

Pangenome visualization / complex loci analysis

Shuangjia Lu, Justin Chu, Christian Fischer, Andrea Guarracino, Erik Garrison

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Population genetic analysis

Silvia Buonaiuto, Andrea Guarracino, Vincenza Colonna, Erik Garrison

Sample selection

Xiaowen Feng, Adam M. Phillippy, Evan E. Eichler, Karen H. Miga, Susan Fairley, Jan O. Korbel, Katherine M. Munson

Pangenome applications: ChIP-seq analysis

Cristian Groza, Guillaume Bourque

Pangenome applications: VNTR analysis

Tsung-Yu Lu, Mark JP Chaisson

Sequencing

Matthew W. Mitchell, Adam M. Phillippy, Evan E. Eichler, Erich D. Jarvis, Karen H. Miga, Robert S. Fulton, Richard E. Green, Miten Jain, Jan O. Korbel, Alexandra P. Lewis, Katherine M. Munson, Hugh E. Olsen, Samuel Sacco

Lab organizer within the HPRC

Matthew W. Mitchell, Mark JP Chaisson, Paul Flicek, Adam M. Phillippy, Evan E. Eichler, David Haussler, Erich D. Jarvis, Karen H. Miga, Ting Wang, Tobias Marschall, Benedict Paten, Richard E. Green, Miten Jain, Valerie A. Schneider

Pangenome applications: RNA-seq analysis

Jonas A. Sibbesen, Jordan M. Eizenga

NHGRI Program Organization

Adam L. Felsenfeld, Baergen I. Schultz, Michael W. Smith, Heidi J. Sofia

Algorithms / Software development

Hanlee P. Ji, HoJoon Lee, Jouni Sirén, Christian Fischer, Santiago Marco-Sola, Glenn Hickey

Ethical, Legal, and Social Implications (ELSI)

Alice B. Popejoy, Karen H. Miga

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Acknowledgements

We would like to acknowledge Shelby Bidwell and other members of the GenBank staff at the National Center for Biotechnology Information (NCBI; NLM/NIH) for their work to release the assemblies into GenBank. Certain commercial equipment, instruments, or materials are identified to specify adequately experimental conditions or reported results. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the equipment, instruments, or materials identified are necessarily the best available for the purpose. Computational infrastructure and support were provided by the Centre for Information and Media Technology at Heinrich Heine University Düsseldorf.

Funding

This work was funded in part by the National Human Genome Research Institute of the National Institutes of Health under award numbers U41HG010972, 1U01HG010973, U41HG007234, 1R01HG011274, R01HG010485, U24HG010262 U01HG010963, U24HG007497. This work was funded in part by the National Institutes of Health under award numbers U01HG010961, OT2OD033761, U24HG011853, R01-HG006677 R35-GM130151, R01HG002385, R01HG010169, HG007497, U01HG010963, U01HG01973, R01HG011649, 5U01HG010971, R01GM123489. U24HG009081, supplement to U24HG009081, R01-HG006677, R35-GM130151, 1ZIAHG200398. The Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health. The work of F.T.-N. And V.A.S was supported by the National Center for Biotechnology Information of the National Library of Medicine (NLM), National Institutes of Health. The USDA National Institute of Food and Agriculture, grant number 2018-67015-28199. The National Science Foundation (NSF), grant IOS-1744309, and NSF PPoSS Award #2118709 (E.G. and P.P.). The Natural Sciences and Engineering Research Council of Canada (NSERC). G.B. is supported by a Canada Research Chair Tier 1 award, a FRQ-S, Distinguished Research Scholar award and by the World Premier International Research Center Initiative (WPI), MEXT, Japan. J.S. was supported by the Carlsberg Foundation. Intramural funding at the National Institute of Standards and Technology. E.E.E., D.H. and E.J. is an investigator of the Howard Hughes Medical Institute. An Oxford Nanopore Research Grant SC20130149 awarded to Mark Akeson, University of California Santa Cruz. The Wellcome Trust, award numbers WT104947/Z/14/Z, WT222155/Z/20/Z and WT108749/Z/15/Z. A Juan de la Cierva fellowship grant (IJC2020-045916-I) funded by MCIN/AEI/ 10.13039/501100011033 and by the European Union NextGenerationEU/PRTR. The Novo Nordisk Foundation (NNF21OC0069089). S.H. acknowledges funding from the Central Innovation Programme (ZIM) for SMEs of the Federal Ministry for Economic Affairs and Energy of Germany. This work was supported by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A537B, 031A533A, 031A538A, 031A533B, 031A535A, 031A537C, 031A534A and 031A532B). The German Federal Ministry of Education and Research (BMBF): 031L0184A. The European Commission, Innovative training network (ITN): 956229. The Taiwan Ministry of Education: Government Scholarship to Study Abroad (GSSA).

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Bibliography

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., & Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74.

Aganezov, S., Yan, S. M., Soto, D. C., Kirsche, M., Zarate, S., Avdeyev, P., Taylor, D. J., Shafin, K., Shumate, A., Xiao, C., Wagner, J., McDaniel, J., Olson, N. D., Sauria, M. E. G., Vollger, M. R., Rhie, A., Meredith, M., Martin, S., Lee, J., … Schatz, M. C. (2022). A complete reference genome improves analysis of human genetic variation. *Science*, *376*(6588), eabl3533.

Alkan, C., Sajjadian, S., & Eichler, E. E. (2011). Limitations of next-generation genome sequence assembly. *Nature Methods*, *8*(1), 61–65.

Altemose, N., Logsdon, G. A., Bzikadze, A. V., Sidhwani, P., Langley, S. A., Caldas, G. V., Hoyt, S. J., Uralsky, L., Ryabov, F. D., Shew, C. J., Sauria, M. E. G., Borchers, M., Gershman, A., Mikheenko, A., Shepelev, V. A., Dvorkina, T., Kunyavskaya, O., Vollger, M. R., Rhie, A., … Miga, K. H. (2022). Complete genomic and epigenetic maps of human centromeres. *Science*, *376*(6588), eabl4178.

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.

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., Warren, W. C., Magrini, V., McGrath, S. D., Li, Y. I., Wilson, R. K., & Eichler, E. E. (2019). Characterizing the Major

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Structural Variant Alleles of the Human Genome. *Cell*, *176*(3), 663–675.e19.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research*, *27*(2), 573–580.

Byrska-Bishop, M., Evani, U. S., Zhao, X., Basile, A. O., Abel, H. J., Regier, A. A., Corvelo, A., Clarke, W. E., Musunuri, R., Nagulapalli, K., Fairley, S., Runnels, A., Winterkorn, L., Lowy, E., The Human Genome Structural Variation Consortium, Flicek, P., Germer, S., Brand, H., Hall, I. M., … Zody, M. C. (2021). High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. In *bioRxiv* (p. 2021.02.06.430068). https://doi.org/10.1101/2021.02.06.430068

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., Fan, X., Wen, J., Handsaker, R. E., Fairley, S., Kronenberg, Z. N., Kong, X., Hormozdiari, F., Lee, D., Wenger, A. M., … Lee, C. (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature Communications*, *10*(1), 1784.

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*, *18*(2), 170–175.

Chen, X., Shen, F., Gonzaludo, N., Malhotra, A., Rogert, C., Taft, R. J., Bentley, D. R., & Eberle, M. A. (2021). Cyrius: accurate CYP2D6 genotyping using whole-genome sequencing data. *The Pharmacogenomics Journal*, *21*(2), 251–261.

Chiang, C., Scott, A. J., Davis, J. R., Tsang, E. K., Li, X., Kim, Y., Hadzic, T., Damani, F. N., Ganel, L., GTEx Consortium, Montgomery, S. B., Battle, A., Conrad, D. F., & Hall, I. M. (2017). The impact of structural variation on human gene expression. *Nature Genetics*, *49*(5), 692–699.

Chin, C.-S., Wagner, J., Zeng, Q., Garrison, E., Garg, S., Fungtammasan, A., Rautiainen, M., Aganezov, S., Kirsche, M., Zarate, S., Schatz, M. C., Xiao, C., Rowell, W. J., Markello, C., Farek, J., Sedlazeck, F. J., Bansal, V., Yoo, B., Miller, N., … Zook, J. M. (2020). A diploid

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

assembly-based benchmark for variants in the major histocompatibility complex. *Nature Communications*, *11*(1), 4794.

Choo, K. H., Vissel, B., Brown, R., Filby, R. G., & Earle, E. (1988). Homologous alpha satellite sequences on human acrocentric chromosomes with selectivity for chromosomes 13, 14 and 21: implications for recombination between nonhomologues and Robertsonian translocations. *Nucleic Acids Research*, *16*(4), 1273–1284.

Choo, K. H., Vissel, B., & Earle, E. (1989). Evolution of alpha-satellite DNA on human acrocentric chromosomes. *Genomics*, *5*(2), 332–344.

Cleary, J. G., Braithwaite, R., Gaastra, K., Hilbush, B. S., Inglis, S., Irvine, S. A., Jackson, A., Littin, R., Rathod, M., Ware, D., Zook, J. M., Trigg, L., & De La Vega, F. M. (2015). Comparing Variant Call Files for performance benchmarking of next-generation sequencing variant calling pipelines. In *bioRxiv*. https://doi.org/10.1101/023754

Collins, R. L., Brand, H., Karczewski, K. J., Zhao, X., Alföldi, J., Francioli, L. C., Khera, A. V., Lowther, C., Gauthier, L. D., Wang, H., Watts, N. A., Solomonson, M., O'Donnell-Luria, A., Baumann, A., Munshi, R., Walker, M., Whelan, C. W., Huang, Y., Brookings, T., … Talkowski, M. E. (2020). A structural variation reference for medical and population genetics. *Nature*, *581*(7809), 444–451.

Comparative genomics: the bacterial pan-genome. (2008). *Current Opinion in Microbiology*, *11*(5), 472–477.

Computational Pan-Genomics Consortium. (2018). Computational pan-genomics: status, promises and challenges. *Briefings in Bioinformatics*, *19*(1), 118–135.

Cunningham, F., Allen, J. E., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Austine-Orimoloye, O., Azov, A. G., Barnes, I., Bennett, R., Berry, A., Bhai, J., Bignell, A., Billis, K., Boddu, S., Brooks, L., Charkhchi, M., Cummins, C., Da Rin Fioretto, L., … Flicek, P. (2022). Ensembl 2022. *Nucleic Acids Research*, *50*(D1), D988–D995.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A.,

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2). https://doi.org/10.1093/gigascience/giab008

Davis, C. A., Hitz, B. C., Sloan, C. A., Chan, E. T., Davidson, J. M., Gabdank, I., Hilton, J. A., Jain, K., Baymuradov, U. K., Narayanan, A. K., Onate, K. C., Graham, K., Miyasato, S. R., Dreszer, T. R., Strattan, J. S., Jolanki, O., Tanaka, F. Y., & Cherry, J. M. (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Research*, *46*(D1), D794–D801.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* , *29*(1), 15–21.

Doerr, D. (2022). *Pangenome-growth: calculate growth statistics for pangenome graphs*. GitHub. https://github.com/marschall-lab/pangenome-growth

Ebert, P., Audano, P. A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., Bonder, M. J., Sulovari, A., Ebler, J., Zhou, W., Serra Mari, R., Yilmaz, F., Zhao, X., Hsieh, P., Lee, J., Kumar, S., Lin, J., Rausch, T., Chen, Y., Ren, J., … Eichler, E. E. (2021). Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*, *372*(6537). https://doi.org/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., 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. (2021). 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,

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

T., Sirén, J., & Garrison, E. (2020). Pangenome Graphs. *Annual Review of Genomics and Human Genetics*, *21*, 139–162.

ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74.

English, A. C., Menon, V. K., Gibbs, R., Metcalf, G. A., & Sedlazeck, F. J. (2022). Truvari: Refined structural variant comparison preserves Allelic diversity. In *bioRxiv*. https://doi.org/10.1101/2022.02.21.481353

Falchi, M., El-Sayed Moustafa, J. S., Takousis, P., Pesce, F., Bonnefond, A., Andersson-Assarsson, J. C., Sudmant, P. H., Dorajoo, R., Al-Shafai, M. N., Bottolo, L., Ozdemir, E., So, H.-C., Davies, R. W., Patrice, A., Dent, R., Mangino, M., Hysi, P. G., Dechaume, A., Huyvaert, M., … Froguel, P. (2014). Low copy number of the salivary amylase gene predisposes to obesity. *Nature Genetics*, *46*(5), 492–497.

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.

Fiddes, I. T., Lodewijk, G. A., Mooring, M., Bosworth, C. M., Ewing, A. D., Mantalas, G. L., Novak, A. M., van den Bout, A., Bishara, A., Rosenkrantz, J. L., Lorig-Roach, R., Field, A. R., Haeussler, M., Russo, L., Bhaduri, A., Nowakowski, T. J., Pollen, A. A., Dougherty, M. L., Nuttle, X., … Haussler, D. (2018). Human-Specific NOTCH2NL Genes Affect Notch Signaling and Cortical Neurogenesis. *Cell*, *173*(6), 1356–1369.e22.

Frankish, A., Diekhans, M., Jungreis, I., Lagarde, J., Loveland, J. E., Mudge, J. M., Sisu, C., Wright, J. C., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Boix, C., Carbonell Sala, S., Cunningham, F., Di Domenico, T., Donaldson, S., Fiddes, I. T., García Girón, C., … Flicek, P. (2021). GENCODE 2021. *Nucleic Acids Research*, *49*(D1), D916–D923.

Gao, Y., Liu, Y., Ma, Y., Liu, B., Wang, Y., & Xing, Y. (2021). abPOA: an SIMD-based C library

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

for fast partial order alignment using adaptive band. *Bioinformatics* , *37*(15), 2209–2211.

Garg, S. (2020). *Pstools: a toolkit for fully phased sequences on chromosome level*. GitHub. https://github.com/shilpagarg/pstools

Garg, S. (2021). Computational methods for chromosome-scale haplotype reconstruction. *Genome Biology*, *22*(1), 101.

Garrison, E. (2021). *Vcfbub: popping bubbles in vg deconstruct VCFs*. GitHub. https://github.com/pangenome/vcfbub

Garrison, E., Guarracino, A., Heumos, S., Villani, F., Bao, Z., Tattini, L., Hagmann, J., Vorbrugg, S., Ashbrook, D. G., Thorell, K., Chen, H., Sudmant, P. H., Liti, G., Colonna, V., & Prins, P. (2022). *The PanGenome Graph Builder*.

Garrison, E., Kronenberg, Z. N., Dawson, E. T., Pedersen, B. S., & Prins, P. (2022). A spectrum of free software tools for processing the VCF variant call format: vcflib, bio-vcf, cyvcf2, hts-nim and slivar. *PLoS Computational Biology*, *18*(5), e1009123.

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.

github.com/vgteam/vg_wdl/GiraffeDeepVariantLite. (2022). Zenodo. https://doi.org/10.5281/ZENODO.6655968

Groza, C., & Bourque, G. (2022). *Epigenomic analysis on the HPRC genome graphs* [Data set]. Zenodo. https://doi.org/10.5281/ZENODO.6564396

Groza, C., Chen, X., Pacis, A., Simon, M.-M., Pramatarova, A., Aracena, K. A., Pastinen, T., Barreiro, L. B., & Bourque, G. (2022). Genome graphs detect human polymorphisms in active epigenomic state during influenza infection. In *bioRxiv* (p. 2021.09.29.462206).

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

https://doi.org/10.1101/2021.09.29.462206

Groza, C., Kwan, T., Soranzo, N., Pastinen, T., & Bourque, G. (2020). Personalized and graph genomes reveal missing signal in epigenomic data. *Genome Biology*, *21*(1), 124.

Grytten, I., Rand, K. D., Nederbragt, A. J., Storvik, G. O., Glad, I. K., & Sandve, G. K. (2019). Graph Peak Caller: Calling ChIP-seq peaks on graph-based reference genomes. *PLoS Computational Biology*, *15*(2), e1006731.

Guarracino, A., Buonaito, S., Rhie, A., Potapova, T., Gerton, J., Colonna, V., Phillippy, A., Human Pangenome Reference Consortium, & Garrison, E. (2022). *Chromosome communities in the human pangenome*. Zenodo. https://doi.org/10.5281/ZENODO.6532467

Guarracino, A., Heumos, S., Nahnsen, S., Prins, P., & Garrison, E. (2022). ODGI: understanding pangenome graphs. *Bioinformatics* . https://doi.org/10.1093/bioinformatics/btac308

Handsaker, R. E., Van Doren, V., Berman, J. R., Genovese, G., Kashin, S., Boettger, L. M., & McCarroll, S. A. (2015). Large multiallelic copy number variations in humans. *Nature Genetics*, *47*(3), 296–303.

Heller, D., & Vingron, M. (2019). SVIM: structural variant identification using mapped long reads. *Bioinformatics* , *35*(17), 2907–2915.

Heller, D., & Vingron, M. (2020). SVIM-asm: Structural variant detection from haploid and diploid genome assemblies. *Bioinformatics* . https://doi.org/10.1093/bioinformatics/btaa1034

Hickey, G. (2021). *Hal2vg: convert HAL to vg-compatible sequence graph*. GitHub. https://github.com/ComparativeGenomicsToolkit/hal2vg

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., Li, H., & Paten, B. (2022). *The Minigraph-Cactus Pangenome Construction Pipeline*.

Hickey, G., Monlong, J., Li, H., & Paten, B. (in preparation). *Pangenome Graph Construction*

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

using Whole-Genome Alignment.

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.

International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921.

Jain, C., Rhie, A., Hansen, N. F., Koren, S., & Phillippy, A. M. (2022). Long-read mapping to repetitive reference sequences using Winnowmap2. *Nature Methods*. https://doi.org/10.1038/s41592-022-01457-8

Jarvis, E. D., Formenti, G., Rhie, A., Guarracino, A., Yang, C., Wood, J., Tracey, A., Thibaud-Nissen, F., Vollger, M. R., Porubsky, D., Cheng, H., Asri, M., Logsdon, G. A., Carnevali, P., Chaisson, M. J. P., Chin, C.-S., Cody, S., Collins, J., Ebert, P., … Human Pangenome Reference Consortium. (2022). Automated assembly of high-quality diploid human reference genomes. In *bioRxiv* (p. 2022.03.06.483034). https://doi.org/10.1101/2022.03.06.483034

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., … MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, *581*(7809), 434–443.

Kirsche, M., Prabhu, G., Sherman, R., Ni, B., Aganezov, S., & Schatz, M. C. (2021). Jasmine: Population-scale structural variant comparison and analysis. In *bioRxiv*. bioRxiv. https://doi.org/10.1101/2021.05.27.445886

Kolesnikov, A., Goel, S., Nattestad, M., Yun, T., Baid, G., Yang, H., McLean, C. Y., Chang, P.-C., & Carroll, A. (2021). DeepTrio: Variant Calling in Families Using Deep Learning. In *bioRxiv* (p. 2021.04.05.438434). https://doi.org/10.1101/2021.04.05.438434

Krusche, P., Trigg, L., Boutros, P. C., Mason, C. E., De La Vega, F. M., Moore, B. L., Gonzalez-

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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., & Global Alliance for Genomics and Health Benchmarking Team. (2019). Best practices for benchmarking germline small-variant calls in human genomes. *Nature Biotechnology*, *37*(5), 555–560.

Larson, D. E., Abel, H. J., Chiang, C., Badve, A., Das, I., Eldred, J. M., Layer, R. M., & Hall, I. M. (2019). svtools: population-scale analysis of structural variation. *Bioinformatics* , *35*(22), 4782–4787.

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

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. In *arXiv [q-bio.GN]*. arXiv. https://doi.org/10.48550/ARXIV.1303.3997

Li, H. (2018a). *Seqtk: a toolkit for processing sequences in FASTA/Q formats*. GitHub. https://github.com/lh3/seqtk

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

Li, H. (2019a). *ETRF: exact tandem repeat finder*. GitHub. https://github.com/lh3/etrf

Li, H. (2019b). *SDUST: symmetric DUST for finding low-complexity regions in DNA sequences*. GitHub. https://github.com/lh3/sdust

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

Li, H. (2020). *Yak: yet another k-mer analyzer*. GitHub. https://github.com/lh3/yak

Li, H. (2021a). *Gfatools: tools for manipulating sequence graphs in the GFA and rGFA formats*. GitHub. https://github.com/lh3/gfatools

Li, H. (2021b). New strategies to improve minimap2 alignment accuracy. *Bioinformatics* . https://doi.org/10.1093/bioinformatics/btab705

Li, H., Bloom, J. M., Farjoun, Y., Fleharty, M., Gauthier, L., Neale, B., & MacArthur, D. (2018). A

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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), 265.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* , *25*(16), 2078–2079.

Lin, M. F. (2021). *Gfabase: GFA insert into GenomicSQLite*. GitHub. https://github.com/mlin/gfabase

Logsdon, G. A., Vollger, M. R., Hsieh, P., Mao, Y., Liskovykh, M. A., Koren, S., Nurk, S., Mercuri, L., Dishuck, P. C., Rhie, A., de Lima, L. G., Dvorkina, T., Porubsky, D., Harvey, W. T., Mikheenko, A., Bzikadze, A. V., Kremitzki, M., Graves-Lindsay, T. A., Jain, C., … Eichler, E. E. (2021). The structure, function and evolution of a complete human chromosome 8. *Nature*, *593*(7857), 101–107.

Lu, T.-Y., Human Genome Structural Variation Consortium, & Chaisson, M. J. P. (2021). Profiling variable-number tandem repeat variation across populations using repeat-pangenome graphs. *Nature Communications*, *12*(1), 4250.

MacArthur, D. G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J. K., Montgomery, S. B., Albers, C. A., Zhang, Z. D., Conrad, D. F., Lunter, G., Zheng, H., Ayub, Q., DePristo, M. A., Banks, E., Hu, M., … Tyler-Smith, C. (2012). A systematic survey of loss-of-function variants in human protein-coding genes. *Science*, *335*(6070), 823–828.

Marco-Sola, S., Eizenga, J. M., Guarracino, A., Paten, B., Garrison, E., & Moreto, M. (2022). Optimal gap-affine alignment in *O*(*s*) space. In *bioRxiv*. https://doi.org/10.1101/2022.04.14.488380

Martin, M., Patterson, M., Garg, S., O Fischer, S., Pisanti, N., Klau, G. W., Schöenhuth, A., &

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Marschall, T. (2016). WhatsHap: fast and accurate read-based phasing. In *bioRxiv*. bioRxiv. https://doi.org/10.1101/085050

Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D., & Gurevich, A. (2018). Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* , *34*(13), i142–i150.

Miller, N. A., Farrow, E. G., Gibson, M., Willig, L. K., Twist, G., Yoo, B., Marrs, T., Corder, S., Krivohlavek, L., Walter, A., Petrikin, J. E., Saunders, C. J., Thiffault, I., Soden, S. E., Smith, L. D., Dinwiddie, D. L., Herd, S., Cakici, J. A., Catreux, S., … Kingsmore, S. F. (2015). A 26-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases. *Genome Medicine*, *7*, 100.

Mohajeri, K., Cantsilieris, S., Huddleston, J., Nelson, B. J., Coe, B. P., Campbell, C. D., Baker, C., Harshman, L., Munson, K. M., Kronenberg, Z. N., Kremitzki, M., Raja, A., Catacchio, C. R., Graves, T. A., Wilson, R. K., Ventura, M., & Eichler, E. E. (2016). Interchromosomal core duplicons drive both evolutionary instability and disease susceptibility of the Chromosome 8p23.1 region. *Genome Research*, *26*(11), 1453–1467.

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.

Numanagic, I., Gökkaya, A. S., Zhang, L., Berger, B., Alkan, C., & Hach, F. (2018). Fast characterization of segmental duplications in genome assemblies. *Bioinformatics* , *34*(17), i706–i714.

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.

Olson, N. D., Wagner, J., McDaniel, J., Stephens, S. H., Westreich, S. T., Prasanna, A. G.,

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Johanson, E., Boja, E., Maier, E. J., Serang, O., Jáspez, D., Lorenzo-Salazar, J. M., Muñoz-Barrera, A., Rubio-Rodríguez, L. A., Flores, C., Kyriakidis, K., Malousi, A., Shafin, K., Pesout, T., … Zook, J. M. (2022). PrecisionFDA Truth Challenge V2: Calling variants from short and long reads in difficult-to-map regions. *Cell Genomics*, *2*(5). https://doi.org/10.1016/j.xgen.2022.100129

Pacific Biosciences. (2021). *PBSV: a suite of tools to call and analyze structural variants in diploid genomes from PacBio SMRT reads*. GitHub. https://github.com/PacificBiosciences/pbsv

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., Novak, A. M., Eizenga, J. M., & Garrison, E. (2017). Genome graphs and the evolution of genome inference. *Genome Research*, *27*(5), 665–676.

Pedersen, B. S., Brown, J. M., Dashnow, H., Wallace, A. D., Velinder, M., Tristani-Firouzi, M., Schiffman, J. D., Tvrdik, T., Mao, R., Best, D. H., Bayrak-Toydemir, P., & Quinlan, A. R. (2021). Effective variant filtering and expected candidate variant yield in studies of rare human disease. *NPJ Genomic Medicine*, *6*(1), 60.

Pedersen, B. S., & Quinlan, A. R. (2018). Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* , *34*(5), 867–868.

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.

Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Van der Auwera, G. A., Kling, D. E., Gauthier, L. D., Levy-Moonshine, A., Roazen, D., Shakir, K., Thibault, J., Chandran, S., Whelan, C., Lek, M., Gabriel, S., Daly, M. J., Neale, B., MacArthur, D. G., &

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Banks, E. (2017). Scaling accurate genetic variant discovery to tens of thousands of samples. In *bioRxiv*. bioRxiv. https://doi.org/10.1101/201178

Porubsky, D., Ebert, P., Audano, P. A., Vollger, M. R., Harvey, W. T., Marijon, P., Ebler, J., Munson, K. M., Sorensen, M., Sulovari, A., Haukness, M., Ghareghani, M., Human Genome Structural Variation Consortium, Lansdorp, P. M., Paten, B., Devine, S. E., Sanders, A. D., Lee, C., Chaisson, M. J. P., … Marschall, T. (2021). Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nature Biotechnology*, *39*(3), 302–308.

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* , *26*(6), 841–842.

Rahman, A., & Pachter, L. (2013). CGAL: computing genome assembly likelihoods. *Genome Biology*, *14*(1), R8.

Rautiainen, M., & Marschall, T. (2020). GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome Biology*, *21*(1), 253.

Rhie, A., Walenz, B. P., Koren, S., & Phillippy, A. M. (2020). Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology*, *21*(1), 245.

Robinson, J. T., Thorvaldsdóttir, H., Wenger, A. M., Zehir, A., & Mesirov, J. P. (2017). Variant Review with the Integrative Genomics Viewer. *Cancer Research*, *77*(21), e31–e34.

Schatz, M. C., Philippakis, A. A., Afgan, E., Banks, E., Carey, V. J., Carroll, R. J., Culotti, A., Ellrott, K., Goecks, J., Grossman, R. L., Hall, I. M., Hansen, K. D., Lawson, J., Leek, J. T., Luria, A. O., Mosher, S., Morgan, M., Nekrutenko, A., O'Connor, B. D., … Wuichet, K. (2022). Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space. *Cell Genomics*, *2*(1). https://doi.org/10.1016/j.xgen.2021.100085

Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., &

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Schatz, M. C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*, *15*(6), 461–468.

Shafin, K., Pesout, T., Lorig-Roach, R., Haukness, M., Olsen, H. E., Bosworth, C., Armstrong, J., Tigyi, K., Maurer, N., Koren, S., Sedlazeck, F. J., Marschall, T., Mayes, S., Costa, V., Zook, J. M., Liu, K. J., Kilburn, D., Sorensen, M., Munson, K. M., … Paten, B. (2020). Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nature Biotechnology*, *38*(9), 1044–1053.

Shumate, A., & Salzberg, S. L. (2020). Liftoff: accurate mapping of gene annotations. *Bioinformatics* . https://doi.org/10.1093/bioinformatics/btaa1016

Sibbesen, J. A., Eizenga, J. M., Novak, A. M., Sirén, J., Chang, X., Garrison, E., & Paten, B. (2021). Haplotype-aware pantranscriptome analyses using spliced pangenome graphs. In *bioRxiv* (p. 2021.03.26.437240). https://doi.org/10.1101/2021.03.26.437240

Sim, S. (2021). *HiFiAdapterFilt: remove CCS reads with remnant PacBio adapter sequences and convert outputs to a compressed .fastq (.fastq.gz)*. GitHub. https://github.com/sheinasim/HiFiAdapterFilt

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.

Sirén, J., & Paten, B. (2022). GBZ File Format for Pangenome Graphs. In *Bioinformatics*.

Slater, G. S. C., & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, *6*, 31.

Smit, AFA, Hubley, R & Green, P. (2013-2015). *RepeatMasker Home Page* (Version RepeatMasker Open-4.0) [Computer software]. http://www.repeatmasker.org/

Sudmant, P. H., Kitzman, J. O., Antonacci, F., Alkan, C., Malig, M., Tsalenko, A., Sampas, N.,

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Bruhn, L., Shendure, J., 1000 Genomes Project, & Eichler, E. E. (2010). Diversity of human copy number variation and multicopy genes. *Science*, *330*(6004), 641–646.

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M. H.-Y., Konkel, M. K., Malhotra, A., Stütz, A. M., Shi, X., Casale, F. P., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., … Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, *526*(7571), 75–81.

Tilgner, H., Grubert, F., Sharon, D., & Snyder, M. P. (2014). Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(27), 9869–9874.

VecScreen: Screen for Vector Contamination. (n.d.). Retrieved June 3, 2022, from https://www.ncbi.nlm.nih.gov/tools/vecscreen/

Wagner, J., Olson, N. D., Harris, L., Khan, Z., Farek, J., Mahmoud, M., Stankovic, A., Kovacevic, V., Yoo, B., Miller, N., Rosenfeld, J. A., Ni, B., Zarate, S., Kirsche, M., Aganezov, S., Schatz, M. C., Narzisi, G., Byrska-Bishop, M., Clarke, W., … Zook, J. M. (2022). Benchmarking challenging small variants with linked and long reads. *Cell Genomics*, *2*(5), 100128.

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.

Wang, T., Antonacci-Fulton, L., Howe, K., Lawson, H. A., Lucas, J. K., Phillippy, A. M., Popejoy, A. B., Asri, M., Carson, C., Chaisson, M. J. P., Chang, X., Cook-Deegan, R., Felsenfeld, A. L., Fulton, R. S., Garrison, E. P., Garrison, N. A., Graves-Lindsay, T. A., Ji, H., Kenny, E. E., … Human Pangenome Reference Consortium. (2022). The Human Pangenome Project: a global resource to map genomic diversity. *Nature*, *604*(7906), 437–446.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.09.499321; this version posted July 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P.-C., Hall, R. J., Concepcion, G. T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N. D., Töpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C.-S., Phillippy, A. M., Schatz, M. C., Myers, G., DePristo, M. A., … Hunkapiller, M. W. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, *37*(10), 1155–1162.

Wick, R. R., Schultz, M. B., Zobel, J., & Holt, K. E. (2015). Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* , *31*(20), 3350–3352.

Wilkins, D. (2022). *Gggenes: a ggplot2 extension for drawing gene arrow maps*. GitHub. https://github.com/wilkox/gggenes

Zhao, X., Collins, R. L., Lee, W.-P., Weber, A. M., Jun, Y., Zhu, Q., Weisburd, B., Huang, Y., Audano, P. A., Wang, H., Walker, M., Lowther, C., Fu, J., Gerstein, M. B., Devine, S. E., Marschall, T., Korbel, J. O., Eichler, E. E., Chaisson, M. J. P., … Talkowski, M. E. (2021). Expectations and blind spots for structural variation detection from long-read assemblies and short-read genome sequencing technologies. *American Journal of Human Genetics*. https://doi.org/10.1016/j.ajhg.2021.03.014

Zook, J. (2021). *Genome In A Bottle - v3.0 Genome Stratifications* [Data set]. National Institute of Standards and Technology. https://doi.org/10.18434/mds2-2499

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 [Review of *Extensive sequencing of seven human genomes to characterize benchmark reference materials*]. *Scientific Data*, *3*, 160025.