Leveraging Human Pangenome for Improved Somatic Variant Detection

Qichen Fu[1,8], Zilan Xin[1,8], Benpeng Miao[1], Wenjin Zhang[1], Nahyun Kong[1], Zitian Tang[1], Andrew Ruttenberg[1], Derek Albracht[2], Edward A. Belter Jr[2], John E. Garza[2], Chad Tomlinson[2], Elvisa Mehinovic[1], Jiawei Shen[1,3], Xiaoyu Zhuo[1], Shihua Dong[1], Benjamin K. Johnson[4], Mary F. Majewski[4], Theron Palmer[4], H. Josh Jang[5], Yuchen Cheng[1], Zefan Li[1], Heather A. Lawson[1], Tina Lindsay[2], Daofeng Li[1,2], Robert Fulton[2], Hui Shen[4], Sheng Chih Jin[1,6], SMaHT Network Assembly/Pangenome Working Group, Juan F. Macias-Velasco[1,2*], Ting Wang[1,2,7*]

[1]Department of Genetics, Washington University School of Medicine, St. Louis, MO 63110, USA
[2]McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO 63110, USA
[3]Department of Computer Science, Washington University in St. Louis, St. Louis, MO 63110, USA
[4]Department of Epigenetics, Van Andel Research Institute, Grand Rapids, MI, USA
[5]Department of Cell Biology, Van Andel Research Institute, Grand Rapids, MI, USA
[6]Department of Pediatrics, Washington University School of Medicine, St. Louis, MO 63110, USA.
[7]Edison Family Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, MO 63110, USA
[8]These authors contributed equally
[*]Correspondence: juanfmacias@wustl.edu and twang@wustl.edu

Abstract

Somatic variant detection is technically challenging due to low variant allele fractions, the confounding presence of germline variation, and reference bias. Linear references such as GRCh38 miss sample-specific variation, causing misalignments and incorrect variant calls. Although telomere-to-telomere donor-specific assemblies (DSAs) accurately represent individual genomes, their application is limited by cost and technical barriers. Alternatively, the graph-based human pangenome provides a scalable framework to improve read alignment and perform genome inference. Here, we benchmarked somatic variant detection using GRCh38, graph-based pangenomes, and pangenome-inferred DSAs with a HapMap mixture dataset and the COLO829 melanoma cell line. Pangenome-guided alignment improves read mapping and somatic variant calling accuracy. Furthermore, personalized pangenomes partially reconstruct donor-specific genomic content, improving accuracy, reducing germline contamination, and enabling detection of events in loci absent or poorly represented in GRCh38. These findings demonstrate that graph-based and personalized pangenomes are effective strategies for enhancing somatic variant detection compared with GRCh38.

Introduction

Somatic variants arise after fertilization and accumulate throughout an individual's lifetime.[1] They are primarily caused by errors in DNA replication or exposure to exogenous DNA-damaging agents, such as ultraviolet radiation, chemicals, X-rays, and tobacco smoke.[2–5] If left unrepaired or not being repaired correctly, these DNA lesions can result in base substitutions, small insertions or deletions, or large structural changes.[3]

While most somatic mutations are silent with minimal phenotypic consequences, a subset contributes to cancer, aging, and neurological disorders.[6] However, the exact impact of somatic mosaicism on biological function remains poorly understood.[2,4,5] The Somatic Mosaicism Across Human Tissues (SMaHT) network aims to shed light on how these events arise and how they impact human health and disease.[2] This will be achieved by generating a comprehensive and reliable catalogue of somatic variation across 19 non-diseased tissue types from 150 donors.[2]

A major barrier for this endeavor is the technical challenge of somatic variant detection. Unlike germline variants, somatic mutations typically occur in a small fraction of cells and exhibit considerable heterogeneity across cells, tissues, and individuals. Low variant allele fractions (VAFs), combined with sequencing errors and artifacts introduced during library preparation, make accurate detection particularly difficult. Moreover, when alignment-based methods are used, the accuracy of somatic variant detection relies on the reference genome. GRCh38 is a mosaic haploid genome composed of a small number of individuals.[7,8] It remains incomplete, with 6.7% of primary chromosome scaffolds unresolved, and its ancestry representation is limited, comprising approximately 57% European and 37% African ancestry.[7,8] Traditional somatic variant detection methods based on GRCh38 rely on either paired normal-tumor samples or panel-of-normal strategies to filter out germline variants. These approaches are still vulnerable to biases introduced using an under-representative reference genome.[7,9,10] The genetic distance between GRCh38 and individual genomes creates reference biases that affect read mapping rate and accuracy, which leads to increased incorrect variant calls. For somatic variant calling, reference biases increase germline contamination in the variant call set, leading to higher chances of misclassification between germline and somatic variants[10]. Moreover, somatic variants located in regions absent from GRCh38 are undetectable when using GRCh38 as the reference.[7]

Because no single linear reference genome can adequately capture human genomic diversity, alternative strategies are needed. While high-quality, donor-specific assemblies (DSAs) are optimal references for somatic variant detection, their high cost limits broad adoption. To overcome these limitations, the Human Pangenome Reference Consortium (HPRC) proposed a shift from a fixed linear reference to a pangenome: a genome-graph composed of a collection of individuals that are representative of most of the common genetic variation from diverse populations.[11] The HPRC is assembling genomes from 350 individuals of diverse ancestries, and in 2023, released a draft human pangenome reference consisting of 94 assemblies from 47 samples and a genome graph incorporating 90 haplotypes, anchored by both GRCh38 and T2T-CHM13.[12] This resource has already demonstrated improved detection of single nucleotide variants (SNVs) in short-read sequencing data.[12]

In this study, we presented two strategies to apply pangenome to somatic variant detection: 1) alignment to the pangenome itself 2) use of a pangenome to approximate the germline assembly of the individual to be analyzed. We benchmarked somatic variant calling across three reference types: GRCh38, pangenome graphs, and pangenome-inferred DSAs. Benchmarking was performed using two experimental systems: a HapMap cell line mixture (Figure 1A) and the COLO829 melanoma cell line (Figure 1B). SMaHT network designed and created the HapMap cell line mixture—a mixture of cell lines in which HG005 contributes 83.5% of the mixture (corresponding to germline variants) and five other donors contribute 0.5-10% of the mixture (corresponding to somatic variants) (Figure 1A).[13] The COLO829 melanoma cell line and its matched normal B lymphoblast cell line (COLO829BL) were also used for benchmarking (Figure

1B). We utilize short-read and long-read whole-genome sequencing (WGS) data of two experimental systems and the HapMap mixture benchmarking variant sets[14] to compare alignment metrics, precision/recall of somatic variant calling (SNVs, indels, and SVs), and mutational signatures (single-base substitution (SBS) & double-base substitution (DBS)) using different references (Figure 1C and 1D).

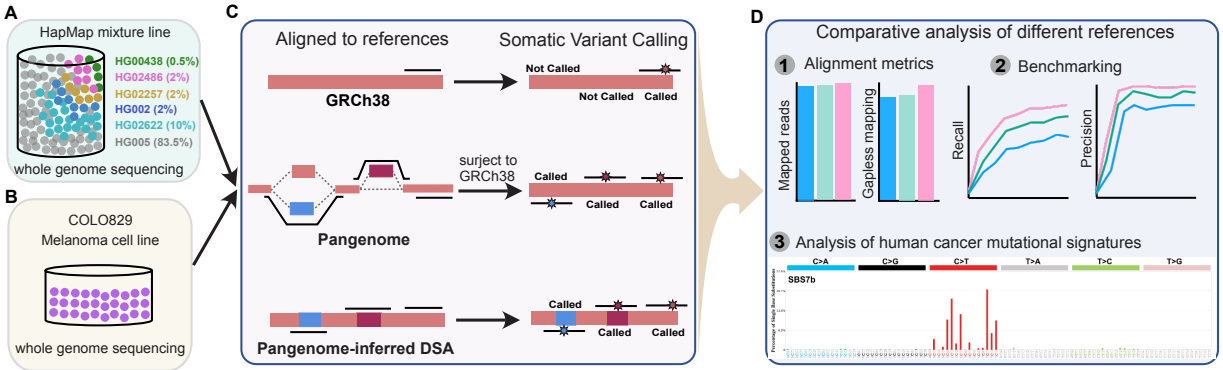

Figure 1. Schematic workflow of the benchmark study. A. Benchmark datasets, including the HapMap mixture (500x Illumina short reads, 100x PacBio HiFi long reads), which was generated by pooling six HapMap cell lines composed of 83.5% HG005 representing the germline background and five additional HapMap donors contributing at 0.5–10% each to simulate somatic variation across a spectrum of VAFs, **B.** and the COLO829 melanoma cell line with its paired normal control (COLO829BL) providing a cancer-derived context. **C.** Overview of the reference types evaluated in this study, including GRCh38, graph-based pangenomes, and pangenome-inferred DSAs. **D.** Alignment of sequencing reads to each reference type using either linear or graph-based aligners, followed by variant calling of SNVs, small indels, and structural variants with appropriate tools. Benchmarking strategy in which variant calls were compared to mixture-derived benchmarking variant sets or tumor–normal comparisons, with stratification by genomic region difficulty and variant allele fraction. Downstream analyses of alignment metrics, variant calling precision and recall, and mutational signature profiling to assess the impact of reference choice on somatic variant detection accuracy.

Our analyses showed that alignment to the pangenome improves alignment quality for short-read and long-read sequencing and subsequently increases the accuracy of somatic variant calling. Pangenome personalization can partially approximate a DSA, which further improves variant calling accuracy, enables better classification between somatic and germline variants, and allows for more comprehensive somatic variant profiling.

Results

Pangenome represents genetic diversity and improves the alignment quality for short and long reads

To benchmark pangenome-based somatic variant detection against linear reference (GRCh38), we utilized HapMap cell line mixture composed of 83.5% HG005 (Male; Chinese) and 10% of HG02622 (Female; Gambian in Western Division, Mandinka), 2% each of HG002 (Male; Ashkenazim Jewish), HG02257 (Female; African Caribbean in Barbados), and HG02486 (Male; African Caribbean in Barbados), and 0.5% of HG00438 (Female; Chinese Han in the South) (Figure 1A)[12]. The variants present in HG005 but absent in any other five individuals are

126 considered as germline variants, whereas the variants present in any of the five individuals but
127 absent in HG005 are considered as somatic variants.
128
129 To evaluate whether the pangenome can represent HapMap germline and somatic samples without
130 their assemblies existing in the graph, we first generated two pangenome graphs using Minigraph-
131 Cactus[15] and compared their nodes, edges, and degree of edges with the HPRC v1.1 pangenome[12].
132 The HPRC v1.1 pangenome (Graph v1 in Figure 2A), constructed by Minigraph-Cactus[15], contains
133 GRCh38, CHM13, and 44 diploid assemblies from diverse genetic backgrounds (including four of
134 the six HapMap samples: HG00438, HG02257, HG02486, HG02622, and excluding HG002 and
135 HG005). The v2 graph contains all HPRC Release 1 assemblies, including HG005, but excluding
136 the five somatic samples. The v3 graph excludes all six HapMap samples (Figure 2A). Three
137 pangenome graphs show a highly similar number of nodes and edges (Figure 2B). The degree (the
138 number of connections to the node) distributions of nodes confirm that the overall graph
139 complexity properties are comparable across versions (Figure 2C). The majority of variants in
140 HapMap samples are captured by other individuals in the v3 pangenome graph, indicating the
141 pangenome graph without HapMap samples can be recognized as a representative reference
142 (Figure S1).
143
144 To determine whether alignment to the pangenome can improve alignment quality, 500x Illumina
145 short-read and 100x PacBio HiFi long-read were aligned to three pangenome graphs (clipped and
146 allele filtered) using vg giraffe[16] (Figure 2D). All pangenome graphs achieved slightly lower
147 numbers of mapped reads (Figure S2). However, the alignment quality was substantially improved
148 with a notable and comparable increase in the number of properly paired, gapless, and perfectly
149 mapped reads for both long and short reads compared with GRCh38 (Figure 2E and S2). For short
150 reads, the gapless and perfectly mapped reads increase by 1.76% and 12.46% respectively (Figure
151 2E). For long reads, the gapless and perfectly mapped reads increase by 425.07% and 966.27%
152 respectively (Figure 2E). To be compatible with linear variant callers, graph alignments were
153 converted to linear (GRCh38) alignments through surjection. Surjection partitions an alignment:
154 the portion aligned to reference nodes remains unchanged, while the portion aligned to non-
155 reference nodes is projected onto nearby reference nodes via local alignment.[12] This process
156 decreases the number of perfectly mapped and gapless reads. However, the pangenome anchors
157 reads in accurate regions and generates more even coverage across the genome compared to
158 GRCh38 (Figure S3). These results indicate that using a pangenome as a more representative
159 reference improves alignment quality compared with GRCh38, substantiating previous results[12,17].

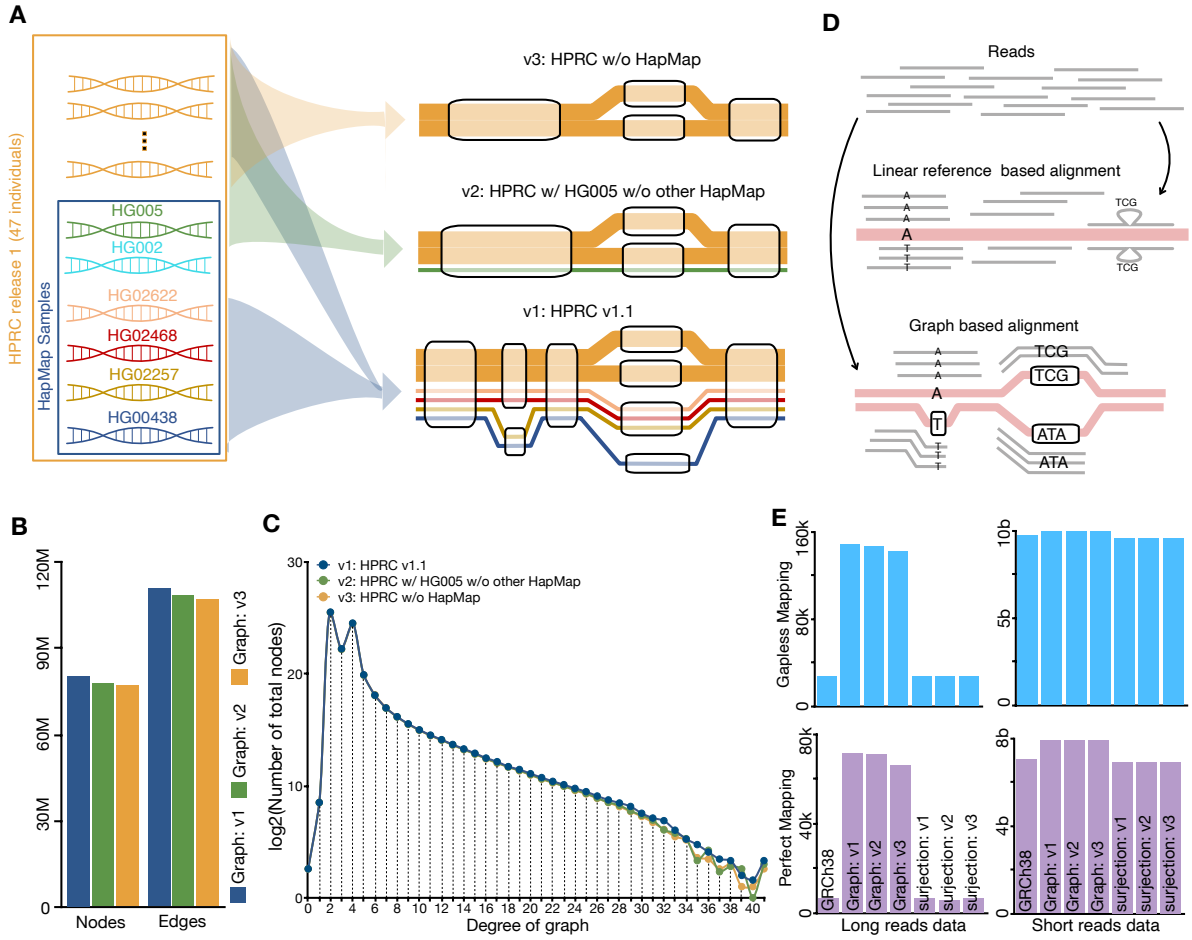

Figure 2. Evaluation of how pangenome composition influences variant representation and alignment quality. Three versions of the draft human pangenome graph were constructed to test whether variants from HapMap mixture donors require explicit inclusion or are already captured through sharing with other assemblies. **A.** Graph v1 includes GRCh38, CHM13, and 44 diploid assemblies (including four HapMap donors). Graph v2 includes all HPRC Release 1 assemblies with HG005 but excludes the five other HapMap donors. Graph v3 excludes all six HapMap donors. **B.** Node and edge counts for the three graphs show highly similar structures, indicating that most variants from the HapMap donors are represented on other assemblies in the pangenome. **C.** Degree distributions confirm that the overall graph complexity properties are comparable across versions. **D.** Illumina short reads (500x) and PacBio HiFi long reads (100x) from the HapMap mixture were aligned to each graph using vg giraffe, resulting in more perfectly mapped and gapless reads than alignment to GRCh38. **E.** Alignment quality metrics comparing graph-based alignments with GRCh38 show that the pangenome improves read placement and alignment score. When graph alignments are surjected back to GRCh38, the proportion of perfectly mapped and gapless reads decreases. In this context, however, the decrease does not indicate poorer alignment quality but instead reflects increased sensitivity to real sequence divergence, such as small indels, that are captured during graph alignment and retained after surjection.

Pangenome improves somatic variant calling accuracy through improved alignment quality

To understand whether the improved alignment quality can improve somatic variant calling accuracy, we compared the variant calling precision and recall from pangenome-guided alignment with GRCh38 alignment. The short-read and long-read WGS data of the HapMap mixture were

aligned to the v3 pangenome graph (clipped and allele filtered) with alignments surjected to GRCh38. Small variants (SNVs & indels <50 bp) were called and filtered from Illumina short read alignments using MuTect2[18]. SVs were called and filtered from PacBio HiFi long reads using Sniffles2[19]. To benchmark the variant calling accuracy, we utilized benchmarking variant sets generated using a genome-graph approach established by Kong et al[14], with GRCh38 as references (Table S1). The pangenome improves SNV calling precision and recall across all VAF bins between 0 and 16.5% (with overall precision improving from 0.9625 to 0.9734 and recall from 0.7145 to 0.7284) (Figure 3A and Table S2). The improvement is greatest in extreme regions, with SNV precision improving from 0.7908 to 0.8506 and recall from 0.4802 to 0.5069 (Figure 3A and Table S2). Moreover, no evident differences in precision or recall are observed across the v1, v2, and v3 pangenome graphs (Figure S4), suggesting that the pangenome can improve variant calling accuracy regardless of whether a DSA is present in the graph.

As an example of how pangenome alignment reduces false negative and false positive somatic variant calls, we highlight a locus at chr12:80,449,114-80,466,689, which contains two repetitive L1PA3 elements (Figure 3B). The somatic variants were detected when reads were aligned to the pangenome but not when aligned to GRCh38. The supporting reads are present in both cases; however, under GRCh38 alignment, they are misplaced to a different genomic region, resulting in apparent false negatives. Incorporating the benchmarking variant set information confirms that the variant resides at the pangenome-supported locus, illustrating how the pangenome improves read localization and enables more accurate somatic variant detection. As indicated in the pangenome deconstructed VCF[12], the two regions have more non-reference sequences, suggesting that the pangenome increases mappability by introducing germline variants that allow aligners to distinguish highly repetitive elements. Graph-based aligners can leverage these non-reference sequences to more effectively localize read placement among highly similar sites.

Alignment to all v1, v2, and v3 pangenome graphs improves somatic indel calling precision (from 0.7410 to 0.7564) and leads to comparable recall between alignment to GRCh38 (0.1867) and pangenome (0.1860) (Figure S5A and Table S2). Precision and recall for SVs are comparable between GRCh38 (0.6640; 0.2024) and pangenome (0.6598; 0.1939) (Figure S5B and Table S2). Moreover, PacBio HiFi long reads span much longer ranges than Illumina short reads, making them less difficult to align accurately on GRCh38. By comparing reads aligned to GRCh38 with those aligned to the pangenome and then surjected to GRCh38, we found that a larger proportion of short reads were mapped to distinct genomic locations between the two approaches compared to long reads (Figure S6), suggesting pangenome benefits less with long read alignment than short read alignment. Together, these results suggest that the pangenome can improve somatic variant calling accuracy through improved alignment quality, especially in challenging genomic regions.

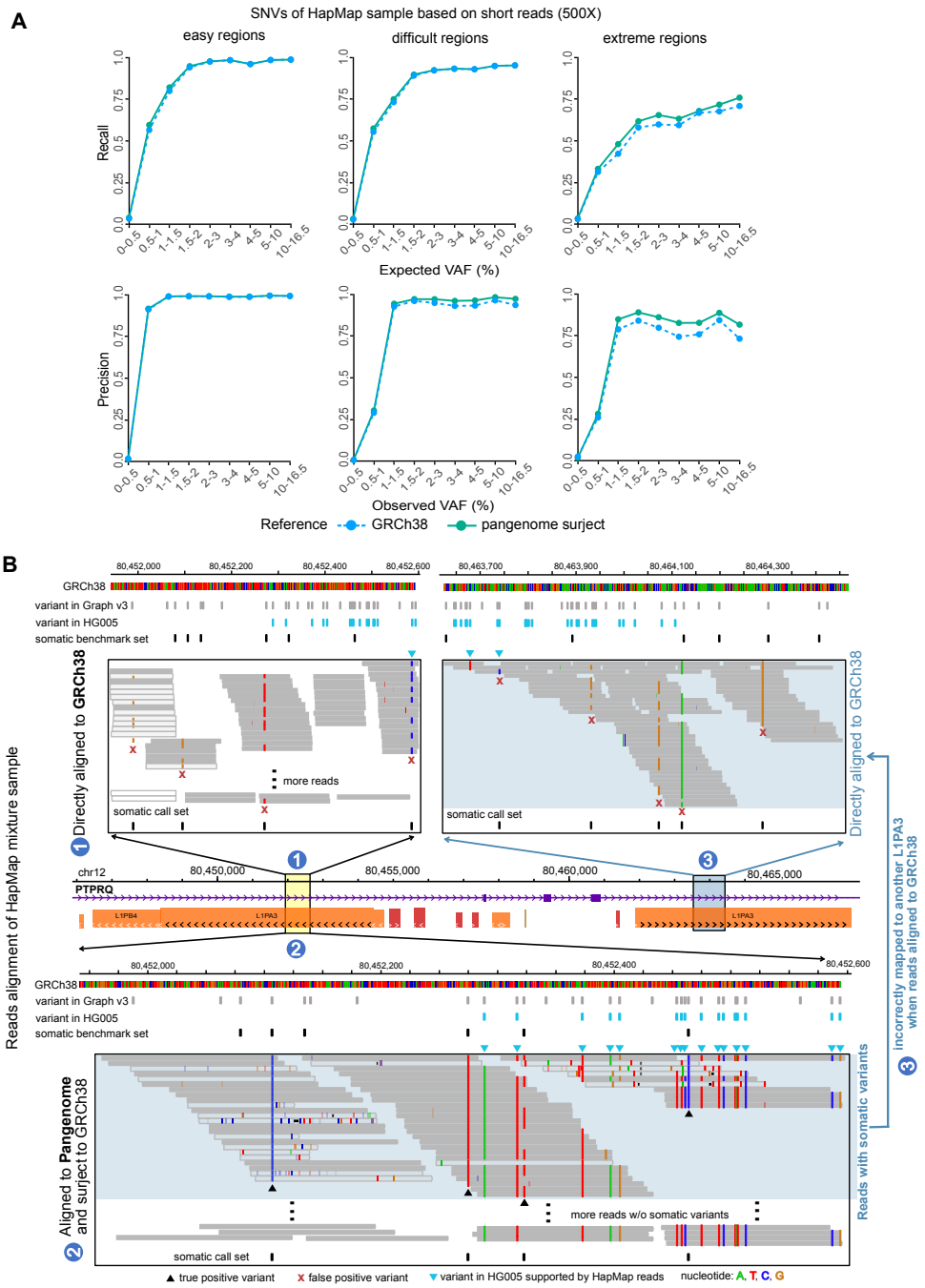

Figure 3. Pangenome improves somatic variant detection in the HapMap mixture. A. Precision and recall for SNVs across VAF bins from 0-16.5%, showing consistent improvements in both measures when using the pangenome, with the largest gains in extreme regions. **B.** Example locus on chromosome 12 where a somatic SNV is detected when aligning reads to the pangenome but not when aligning to GRCh38. Supporting reads are present in both cases; however, under GRCh38 alignment, the reads are misplaced to a different locus, resulting in an apparent false negative. Benchmarking variant set information confirms the variant at the pangenome-supported locus, illustrating how the pangenome improves correct read placement and reduces false negative and false positive calls. Comparison of read placement for the same supporting reads under GRCh38 alignment, showing their misalignment to a different genomic region.

Pangenome enables genome inference, improves somatic variant calling accuracy, and reduces germline contamination

Surjecting reads from the pangenome onto GRCh38 utilizes improved alignments derived from the pangenome representativeness but does not fully take advantage of the completeness and genetic diversity of the pangenome, including sequence that exists in individuals but are missed in GRCh38. Next, we used the pangenome to infer and partially reconstruct the germline assembly.

Variants that are in pangenome graphs but not part of the sample will mislead read mappings. To address this, pangenome personalization has been developed as a strategy to remove irrelevant variants and impute a personalized pangenome subgraph by sampling local haplotypes according to k-mer counts in the individual's sequencing data.[17] This approach further improves read mappings, variant calling, and genotyping.[17] This also allows for leveraging an individual's sequencing data to construct artificial haplotypes that approximate the person's genomic background, thereby enabling the imputation of a DSA.

Here, we used the personalization approach (vg haplotype) to sample the haplotypes that match closest to the k-mer content of the sample from an allele-unfiltered pangenome graph and construct a subgraph containing pseudo-diploid assemblies (hap1 and hap2).[17] The v3 pangenome graph (allele unfiltered, clipped, and excludes all 6 HapMap samples) and k-mer content of HG005 short read WGS were used as an input for pangenome personalization (Methods). To determine how closely inferred DSAs match a high-quality DSA, we aligned GRCh38 and the pangenome-inferred DSAs of HG005 hap1 and hap2 with HG005 diploid assemblies using Minimap2[20]. The pangenome-inferred DSA shows lower genetic divergence with HG005 than GRCh38, with fewer substitutions, insertions, and deletions (Figure 4A). To tolerate haplotype switching between pangenome-inferred DSA hap1 and hap2 in pangenome personalization, we also evaluated the pangenome-inferred DSAs using a graph-based approach (Figure 4B). Both the pangenome-inferred DSAs (hap1 and hap2) and the HG005 diploid assembly were incorporated into a Minigraph-Cactus[15] validation graph. To compare genotypes between the pangenome-inferred assembly and HG005, the validation graph was deconstructed to a VCF with genotype information for four assemblies. Variants in HG005 paternal or maternal assembly were considered captured if they were present in either pangenome-inferred DSA hap1 or hap2. Missing variants were caused by assembly gaps in the HG005 maternal or paternal assembly (Figure 4B). The pangenome-inferred DSAs were found to capture most HG005 variants (Maternal: 84.7% captured, 6.88% missing, 8.38% not captured; Paternal: 86.8% captured, 6.88% missing, 6.33% not captured) (Figure 4C).

Given the improvement in representation from GRCh38 to pangenome-inferred DSA, we next benchmarked somatic variant calling on GRCh38, pangenome-inferred DSAs, HG005 maternal, and HG005 paternal assemblies. Classifying germline and somatic variants becomes challenging when a matched normal is unavailable or inapplicable, such as in normal tissue[21]. The pangenome-inferred DSA filters the majority of HG005 homozygous germline variants at the alignment step (Figure 4D). The VAF distribution of the pangenome-inferred DSA approximates that of the HG005 assemblies (Figure 4D). To benchmark somatic variant calling accuracy, we utilized benchmarking variant sets generated with a genome-graph approach established by Kong et al.[14], with the pangenome-inferred DSAs (hap1 and hap2) and the HG005 maternal and paternal assemblies as references (Figure S7 and Table S1). These sets captured all possible variants across

SNVs, indels, and SVs and were used to compare with respective call sets. The pangenome-inferred DSA hap1 and hap2 achieved precision and recall comparable to the HG005 maternal and paternal assemblies for SNVs, indels, and SVs, and exceeded those of GRCh38 (Figure 4E, S8, and Table S3). The improvement was most pronounced in difficult and extreme regions (Figure 4E, S8, and Table S3). The pangenome-inferred DSA outperforms pangenome-guided alignment in variant calling for SNVs, indels, and SVs (with SNV precision improving from 0.9734 to 0.9764 and recall from 0.7284 to 0.7819 genome-wide; precision from 0.8506 to 0.8871 and recall from 0.5069 to 0.5912 in extreme regions) (Table S2 and S3).

As an example of how pangenome-inferred DSA reduces germline contamination and improves somatic variant calling accuracy, we looked at the same region in Figure 3B. We found that pangenome-inferred DSA correctly identified variants in the first L1PA3 element and filtered germline variants at the alignment step (Figure S9).

Together, these results show that pangenome-guided inference of donor-specific assemblies reduces germline contamination and enhances somatic variant detection to a level approaching that of true DSAs, while avoiding some of the limitations of surjection back to GRCh38.

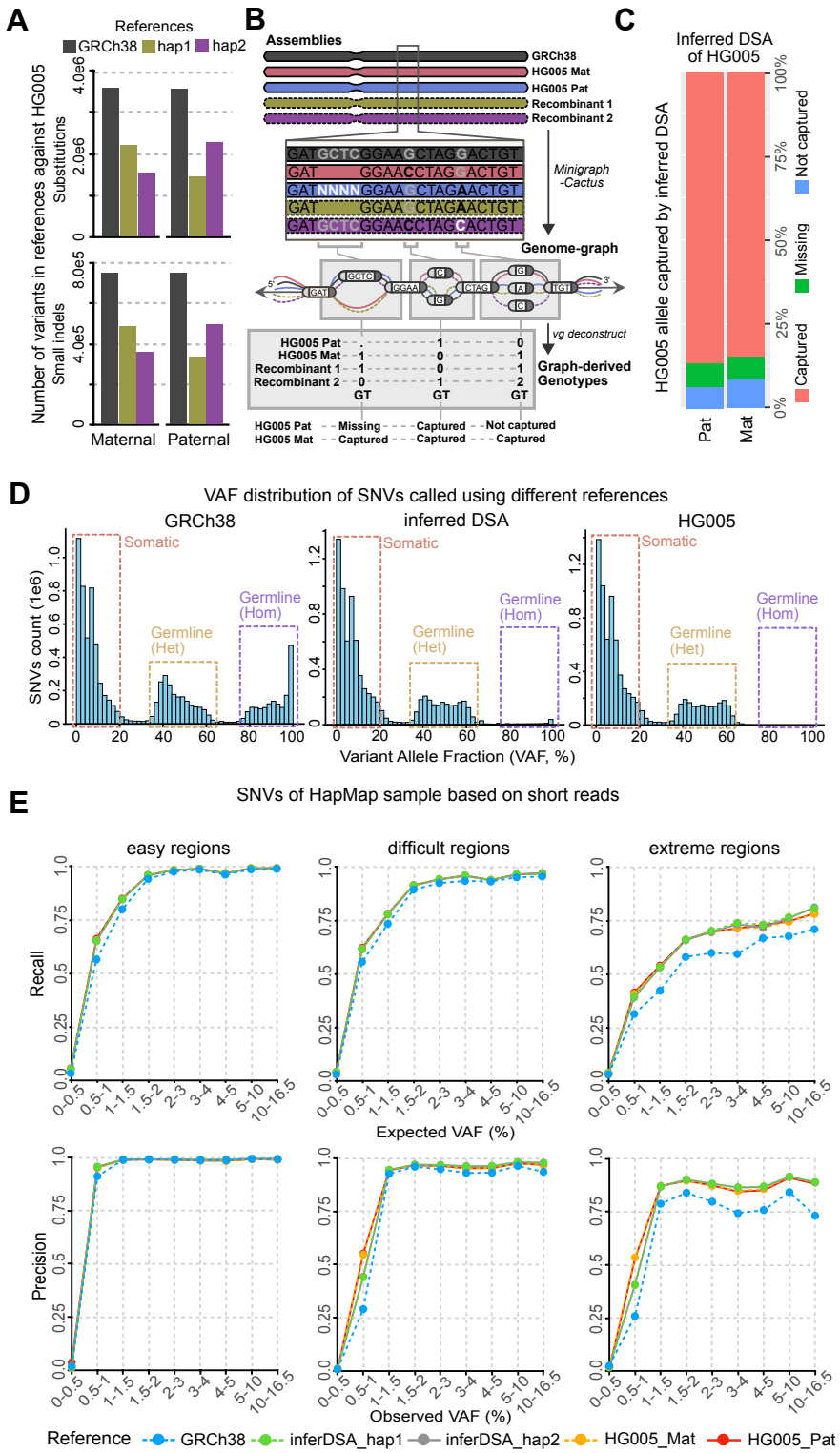

Figure 4. Pangenome-inferred assemblies improve somatic variant detection in the HapMap mixture. Construction of a pangenome-inferred assembly for HG005 using k-mer content from short-read data and haplotype sampling from the v3 pangenome graph, which excluded all 6 HapMap samples. The resulting pseudo-diploid assembly partially reconstructs the HG005 genome. **A.** Comparison of sequence divergence between the inferred HG005 assembly, the true HG005 maternal and paternal assemblies, and GRCh38,

showing that the inferred assembly reduces the number of substitutions relative to GRCh38. **B.** A graph-based method was employed to jointly compare genotypes between the pangenome-inferred assembly and HG005. **C.** Graph-based genotype comparisons show that pangenome-inferred assembly captures approximately 90% of HG005 variants, demonstrating its ability to approximate a donor-specific assembly. **D.** VAF distributions of non-VAF filtered SNV call sets from MuTect2 tumor-only mode for GRCh38, pangenome-inferred DSA (hap1), and HG005 paternal assembly. **E.** Benchmarking of somatic variant calling performance across reference types. The pangenome-inferred DSA achieves precision and recall for SNVs comparable to the HG005 maternal and paternal assemblies and superior to GRCh38, with the largest gains observed in difficult and extreme genomic regions.

Pangenome-inferred DSA allows for more comprehensive somatic SNV profiling for COLO829

Since pangenome-inferred DSA partially reconstructs the germline assembly, we asked whether its call set captures a more comprehensive somatic SNV profile than the GRCh38 call set in the cancer sample. We utilized the PacBio HiFi data of the melanoma cell line COLO829 and its matched normal COLO829BL cell line (B lymphoblast cell line). We profiled the somatic SNVs of COLO829 using GRCh38 or pangenome-inferred DSA of COLO829BL using DeepSomatic[22] tumor-normal mode. The pangenome-inferred DSA was constructed using HPRC v1.1 pangenome (allele unfiltered and clipped) and COLO829BL short-read WGS (20x)[17]. We validated the pangenome-inferred DSA by comparing it with the near T2T DSA of COLO829 generated using Verkko2[23] by the UWSC Genome Characterization Center (GCC)[24]. The pangenome-inferred DSA captures most of the COLO829 germline variants (Maternal: 88.4% captured, 2.73% missing, 8.85% not captured; Paternal: 85.1% captured, 2.73% missing, 12.1% not captured) (Figure 5A). It improves the alignment quality with increased mapped reads and decreased insertion/deletion and supplementary alignments (Figure 5B). The pangenome-inferred DSA detects more somatic SNVs compared to GRCh38, including 4,290 SNVs unique to pangenome-inferred DSA and 930 SNVs unique to GRCh38 (Figure 5C).

Next, to assess the biological relevance of SNVs, we performed mutational signature analysis for shared and unique somatic SNVs from different references. The single-base substitution (SBS) and double-base substitution (DBS) signatures of shared SNVs were enriched in UV light exposure signatures (SBS7a, SBS7b, SBS38, and DBS1) (Figure S10 and S11). Unique SNV calls from pangenome-inferred DSA show higher SNV counts and proportion in UV light exposure SBS signatures (SBS7a, SBS7b, SBS38, and SBS97) than those of the unique SNV calls from GRCh38 (Figure 5C and 5D). DBSs unique to the pangenome-inferred DSA show a five-fold higher count in DBS1 compared with GRCh38-unique DBSs (Figure S10). The increased sensitivity in detecting mutations translated to a slightly higher estimated mutation rate when using pangenome-inferred DSA, both genome-wide and across major genomic features, except for satellite regions. Whereas using GRCh38 the estimated mutation rate in satellite regions was less than half the genome-wide rate, using the pangenome-inferred DSA doubled the estimated mutation rate in these regions (Figure 5E), highlighting that satellite regions are not less susceptible to UV damage and/or the DNA repair machinery is not more efficient[25,26].

Among the 4,290 SNVs unique to the pangenome-inferred DSA, 2,070 of them can be lifted over to GRCh38. We observed that the pangenome-inferred DSA improves regional mappability, increases the MAPQ, and enables somatic variant detection. As an example, we highlighted the

LIMS4 gene (Figure 5F). Across gnomAD[27], GTEx[28], and TCGA[29], LIMS4 displays low population coverage, no detectable germline variants (Figure S12A), and low RNA expression across tissues (Figure S12B) and in melanoma (Figure S12C), yet elevated protein levels in melanoma (Figure S12C). This suggests that variants and RNA levels of the LIMS4 gene were underestimated due to the low mappability in GRCh38. We demonstrated that pangenome-inferred DSA increased the mappability and detected a missense mutation in the exon 1 of LIMS4, which was missed in the GRCh38 call set (Figure 5F). Moreover, the pangenome-inferred DSA aligned a greater number of somatic variant supporting reads, thereby enabling more sensitive and accurate somatic mutation detection (Figure S13).

A total of 2,220 pangenome-inferred DSA unique SNVs could not be lifted over to GRCh38. These SNVs are located in regions missing or poorly represented in GRCh38, including simple repeat expansion (Figure S14A), copy number gain (Figure S14B), and large missing segments, such as satellite regions (Figure 5E). Notably, the pangenome-inferred DSA determined that one haplotype of COLO829BL contains a simple repeat expansion within an intron of GNG10. This discovery is confirmed using COLO829BL DSA generated by UWSC-GCC[24]. A somatic SNV within the repeat expansion was detected using the pangenome-inferred DSA, which was missed by GRCh38 (Figure S14A). The pangenome-inferred DSA can also resolve copy number gains, resulting in improved coverage and more accurate somatic variant detection in extra copies, as illustrated by the example of LINC01410 (Figure S14B). We further summarized the pangenome-inferred DSA unique SNVs that fall within the exon (Table S4). 11 sSNVs were identified as missense mutations, 10 were in cancer genes, and 2 were predicted to be pathogenic by AlphaMissense[30]. 2 missense sSNVs were located in genes previously reported to be associated with melanoma, including ARHGEF5[31], and TMEM87b[32]. 8 missense sSNV were in genes implicated in other cancer types, such as TRIOBP[33], FCGBP[34], and NBPF1[35] (more in discussions and Table S4).

Together, these results highlight that the pangenome can construct a more representative assembly than GRCh38. This more representative assembly allows for more comprehensive and biologically meaningful somatic variant profiling.

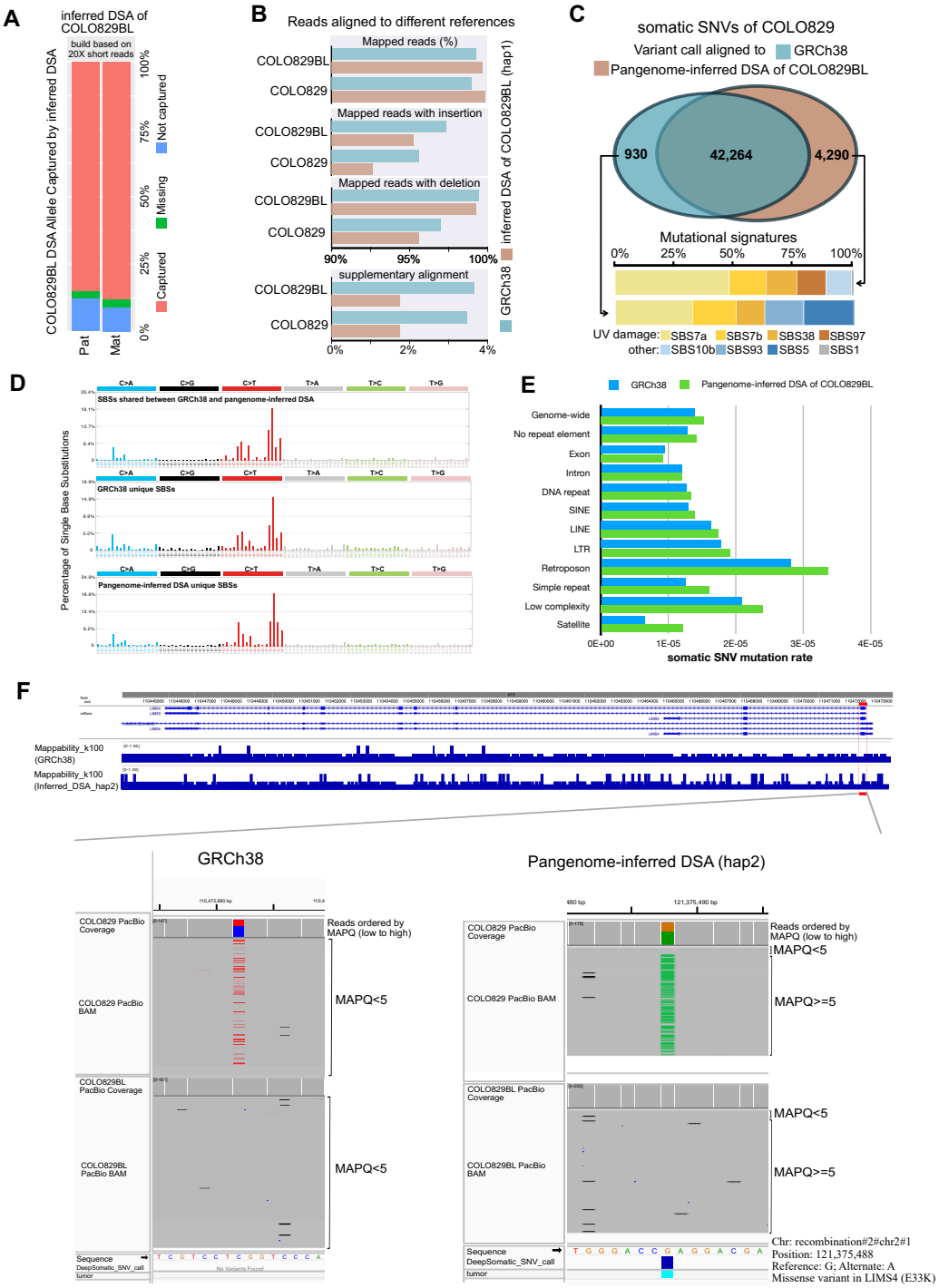

Figure 5. Pangenome-inferred DSA allows for more comprehensive somatic SNV profiling for COLO829. A. Allele capture rate of pangenome-inferred DSA of COLO829BL constructed based on 20x short-read WGS of COLO829BL. **B.** Alignment metrics of COLO829 and COLO829BL PacBio reads aligned to GRCh38 and pangenome-inferred DSA. **C.** The overlap between the SNV call sets using GRCh38 or inferred DSA as reference and the SBS signature decomposition of the unique SNVs. **D.** SBS profiles for shared and unique SNV calls of GRCh38 and pangenome-inferred DSA. **E.** Somatic SNV mutation rate of GRCh38 and pangenome-inferred DSA call set in different genomic features. Exon: total

length of exons from protein-coding genes; Intron: total length of introns from protein-coding genes. **F.** Pangenome-inferred DSA improves the mappability and enables somatic variant detection in the LIMS4 gene.

Discussions

The human pangenome has been developed to capture genetic variation in the human population[12]. It serves as a more complete and representative reference than GRCh38. Here, we presented the first comprehensive benchmark of somatic variant detection using pangenomic methods. We benchmarked pangenome-based somatic variant detection with the GRCh38 approach using two systems – the HapMap mixture and the COLO829 melanoma cell line, designed by the SMaHT Network. The benchmarking variant sets established by the genome-graph approach[14] enable benchmarking of all variant classes, including SNVs, indels, and SVs. We demonstrated that the pangenome can capture most variants of the sample with sample's assembly excluded and allows for improved alignment quality for both short and long read sequencing. This results in improvement in somatic SNV and indel calling, especially in challenging regions. Furthermore, the pangenome can partially reconstruct the germline assembly and improve the accuracy of somatic variant detection to a level comparable with the DSA. This also enables the more comprehensive somatic variant detection for cancer samples such as COLO829.

We demonstrated that the pangenome can improve alignment quality, as evidenced by a substantial increase in gapless and perfectly mapped reads and improved somatic variant calling accuracy. Given that most germline variants of an individual are shared within the human population, the pangenome graph, which represents population-level germline variation, enables more accurate read alignment than GRCh38. To be compatible with linear variant callers, the alignments were converted from the graph to linear (GRCh38) reference through surjection. Surjection partitions an alignment: the portion aligned to reference nodes remains unchanged, while the portion aligned to non-reference nodes is projected onto nearby reference nodes via local alignment.[12] This process guides reads to correct regions and improves somatic SNV and indel calling accuracy, especially in difficult and extreme regions. We demonstrated that, in some repetitive regions, such as transposable elements, the pangenome becomes important for detecting variants correctly (Figure 3B). For optimal SV detection, after converting graph alignments to linear alignments (surjection), reads need to be indel realigned and polished. However, to date, a tool to perform alignment polishing akin to Abra2[36] has not been developed for long reads. Furthermore, we observed that the pangenome provides less benefit for accurately aligning long reads compared to short reads, likely because long reads span much larger genomic regions (Figure S6). These factors may collectively explain the observed comparable performance in SV calling between GRCh38 and the pangenome.

Surjecting reads to GRCh38 limits the advantages of the improved representation and completeness of the pangenome graphs. Alternatively, we used the pangenome graph to impute a personalized subgraph containing a pseudo-diploid assembly that closely approximates the DSAs. Calling somatic variants based on the pseudo-diploid assembly yields notable improvements in alignment and in the accuracy of calling somatic SNVs, indels, and SVs, comparable to those achieved with DSAs and exceeding those from the pangenome surjection approach.

Moreover, our study utilized the HapMap mixture, an artificial system that uses germline variants to simulate somatic variants. This has advantages, as the density of germline variants is much higher than that of true somatic variants, enabling benchmarking across the entire genome, including easy, difficult, and extreme regions. The controlled VAF in the mixture allows for benchmarking at various VAFs.

We demonstrated the advantages of pangenome-inferred DSA for reducing germline contamination in somatic variant detection when a matched normal control is unavailable or inapplicable, e.g., normal tissue. In the HapMap mixture, most homozygous HG005 variants were filtered out by pangenome-inferred DSAs. The pangenome enables filtering germline variants at the alignment step, substantially reducing false positives in the initial call set and alleviating the burden on somatic variant callers of distinguishing somatic variants from a large number of germline variants. In contrast, the widely used Panel of Normals (PoNs) approach filters germline variants after variant calling. PoNs are limited in variant type, as they are widely used for SNVs and small indels but not for structural variants[18]. Pangenome includes assemblies from diverse ancestries, and the filtering is not restricted to any variant type. Our study provides proof of principle that using a representative reference can filter out germline variants during the alignment process, effectively serving as an additional layer of germline and somatic variant classification.

We found that the pangenome-inferred DSA enables more comprehensive profiling of somatic variants in a cancer cell line. Using the pangenome, we identified 4,290 (10.15%) novel SNVs in COLO829. Mutational signature analysis showed that these SNVs are highly enriched for the UV light exposure signature, underscoring their biological relevance in melanoma. We showed that genes with low mappability in GRCh38 are well resolved by the pangenome, emphasizing that a more representative reference—such as a pangenome—is essential for confident read alignment and accurate variant detection for these genes. 51.75% of the SNVs unique to the pangenome-inferred DSA were located in regions that are absent or poorly represented in GRCh38. These SNVs arise from insertions, copy-number gains, and other complex genomic loci, including satellite regions. We showed that GRCh38 underestimates the mutation rate in satellite regions. However, even with improved references, mutation rates in these highly repetitive regions are likely still underestimated. Accurate characterization will require further advances in sequencing technologies and alignment algorithms capable of handling highly repetitive genomic regions.

Many newly discovered somatic SNVs using the pangenome are located in melanoma-associated genes such as GNG10[37], ARHGEF5[31], and TMEM87b[32], as well as in genes implicated in other cancer types, such as LINC01410[38], TRIOBP[33], FCGBP[34], and NBPF1[35]. 11 sSNVs were identified as missense mutations, with 2 predicted to be pathogenic by AlphaMissense[30] (Table S4). 10 of them were located in genes previously reported to be associated with melanoma or other cancer types (Table S4). TMEM87b is the paralog of TMEM87a. TMEM87a was reported as a component of a novel mechanoelectrical transduction pathway and modulates melanoma adhesion and migration[32]. NEB is a prognostic marker in renal cancer based on TCGA[29]. ZSCAN23 is involved in transcription regulation and is a tumor suppressor in pancreatic ductal adenocarcinoma[39]. ARHGEF5 is a key regulator of many cellular events triggered by extracellular signals acting through G protein-coupled receptors and likely contributes to the regulation of cytoskeletal organization[40]. ARHGEF5 promotes epithelial-mesenchymal transition in colorectal cancer[41], and its overexpression is associated with poor prognosis in acute myeloid leukemia[42].

474 LDLRAD4 is involved in the negative regulation of cell migration and epithelial to mesenchymal
475 transition[43]. Decreased LDLRAD4 expression leads to metastasis and poor prognosis in colorectal
476 cancer[44]. ZNF536 is a highly conserved zinc-finger protein that negatively regulates neuronal
477 differentiation by repressing retinoic-acid-induced gene transcription[45]. It has also been linked to
478 poor prognosis in neuroblastoma[46]. Retinoic acid has been reported to affect proliferation in
479 melanoma cell lines[47]. The missense mutation of ZNF536 identified in this study occurs within its
480 C2H2 zinc-finger domain and is predicted to be pathogenic, potentially altering its DNA-binding
481 capability and impacting cell proliferation in COLO829. NBPF1 has been implicated in several
482 types of cancers, including neuroblastoma[35] and adrenocortical carcinoma[48]. TRIOBP regulates
483 actin cytoskeletal organization, cell spreading, and cell contraction. TRIOBP is highly expressed
484 across cancers, and specific isoform expression is associated with gastric, rectal, pancreatic, and
485 brain cancers[49]. FCGBP is involved in the maintenance of the mucosal structure as a gel-like
486 component of the mucosa and is a prognostic or diagnostic marker for multiple cancer types[50].
487 LIMS4 lacks literature support of a cancer gene role, likely due to low mappability in GRCh38,
488 and is predicted to be located in the cytoplasm, focal adhesion, and plasma membrane[43], which
489 could be associated with melanoma metastasis. Collectively, these findings highlight the
490 advantages of using a more representative reference for somatic variant detection and reveal
491 somatic mutations that may be biologically consequential yet missed when relying on GRCh38.
492 This expanded cancer gene mutation catalog may hold new promises in understanding melanoma
493 biology, as well as in developing diagnosis and therapeutics. Pangenome can serve as a great and
494 cost-effective resource for obtaining an individual diploid assembly with reduced reference biases
495 compared with GRCh38.
496
497 Together, these results suggest that pangenome improves read mapping and somatic variant calling
498 accuracy, enables efficient germline variant filtering and more comprehensive somatic variant
499 profiling, and opens new avenues to leveraging pangenomic resources for somatic variant
500 detection.
501
502 Despite these advances, our study has several limitations. First, the pangenome personalization did
503 not fully capture the germline variants of the true DSAs. More sophisticated genome inference
504 methods need to be developed to further minimize the genetic distance between pangenome-
505 inferred DSAs and true DSAs. Second, current somatic variant callers are based on the linear
506 reference genome, preventing full utilization of pangenome sequences and effective filtering of
507 heterozygous germline variants. Graph-based variant callers are needed to address these
508 limitations. Third, the HapMap mixture uses germline variants to simulate somatic variants. The
509 mutational load and pattern may be different from true somatic variants.
510
511 **Methods**
512
513 **HapMap mixture design and generation**
514 To mimic the somatic variants with a broad spectrum of variant allele fractions (VAFs), six well-
515 studied HapMap samples (HG005, HG02622, HG002, HG02257, HG02486, and HG00438) are
516 chosen to construct the cell mixture samples according to certain ratios. The target composition
517 for the pooled sample was 83.5% HG005, 10% HG02622, 2% each of HG002, HG02257, and
518 HG02486, and 0.5% HG00438, among which HG005 is taken as the germline source, while five
519 additional individuals contribute reads simulating artificial somatic variants.

Each cell line was thawed and seeded into T25 flasks containing 10 mL of culture medium composed of 84% RPMI-1640, 15% FBS, and 1% Glutamax. A portion of the cell suspension was used for counting, and the culture volume was adjusted as needed to maintain a viable cell density between 2×10^5 and 5×10^5 cells/mL. Cells were subsequently passed every 3–5 days by dilution based on cell counts until the desired cell number was achieved. Viable cell counts were determined using a Vi-Cell Blu cell counter (Beckman Coulter) with trypan blue staining.

On the day of harvest, cells from each line were pooled into a 250 mL conical tube and mixed thoroughly. Approximately 10 mL of the mixture was transferred to a T25 flask, and an aliquot was taken for cell counting. Sterility was assessed by streaking two sheep blood agar plates, which were incubated at 30 °C and 37 °C for two weeks. Based on the cell counts, the required volume of each cell line was calculated to achieve the target cell number. These volumes were combined in a secondary tube to generate the final pooled suspension. The pooled cells were centrifuged at 228 RCF for 10 minutes, resuspended in 800 mL of cryopreservation medium (65% RPMI-1640, 30% FBS, 5% DMSO), and aliquoted into cryovials at a concentration of 5×10^6 cells per vial. Cryopreservation was performed using a controlled-rate freezer, and vials were stored in liquid nitrogen vapor.

HapMap mixture sequencing data generation

The WashU-VAI Genome Characterization Center produced the 500× short-read WGS data using Illumina NovaSeq X Plus and 100× long-read WGS data using PacBio Revio for HapMap mixture cell lines.

Short-read sequencing

Genomic DNA was quantified using a Qubit Fluorometer. Approximately 600-1,000 ng of DNA was sheared on a Covaris LE220 instrument to generate ~375 bp fragments. Size selection was performed with a 0.8× ratio of AMPure XP beads (Beckman Coulter) to remove fragments shorter than 300 bp. Dual-indexed libraries were prepared using the KAPA Hyper PCR-free Library Preparation Kit (Roche Diagnostics, Cat #7962371001), incorporating full-length custom adapters (IDT) in a UDI/UMI configuration, featuring 10 bp unique dual indexes (UDIs) and a 9 bp unique molecular identifier (UMI) in the i7 position. Library molarity was quantified using the KAPA Library Quantification Kit (Roche Diagnostics). Libraries of 150 bp paired-end reads were sequenced on the NovaSeq X platform. To achieve >500× coverage, four independent libraries were prepared and sequenced, and the resulting data were pooled from 4 library sources.
Library molarity was accurately quantified by qPCR using the KAPA Library Quantification Kit (KAPA Biosystems/Roche) following the manufacturer's instructions, ensuring optimal cluster density for sequencing on the Illumina NovaSeq X Plus platform. Normalized libraries were then sequenced on a NovaSeq X Plus flow cell using a 151×10×10×151 paired-end sequencing protocol, generating whole-genome sequencing coverage exceeding 500×.

Long-read sequencing

PacBio HiFi SMRTbell libraries were prepared based on the Procedure & Checklist – Preparing Whole Genome and Metagenome Libraries Using SMRTbell Prep Kit 3.0 (PacBio). Genomic DNA was initially treated with DNAFluid+ (P/N E07020001) at a speed setting of 40 on the Diagenode Megaruptor 3 to dissociate aggregates and homogenize the sample. The homogenized

DNA was then sheared twice using the Shearing Kit (P/N E07010003) at speeds of 28 and 30, targeting a fragment size mode of ~20 kb. DNA quality and fragment size distribution were assessed using the Qubit High Sensitivity DNA Kit and the Agilent Femto Pulse system (Genomic DNA 165 kb Kit).

Library preparation followed the PacBio protocol and included barcoded adapters from the SMRTbell Barcoded Adapter Plate 3.0 (PacBio, P/N 102-009-200) for multiplexed sequencing. Libraries were size-selected on the Sage PippinHT using the 0.75% Agarose High-Pass 75E Kit (P/N HPE7510) with a lower cutoff between 15,000–17,000 bp. Final library preparation for sequencing was performed according to instructions generated by SMRT Link v13.0, using the PacBio Revio Polymerase Kit (P/N 102-817-600).

Sequencing was conducted on the PacBio Revio system using 30-hour movie collections with an on-plate concentration of 170–200 pM. To achieve ~100x coverage (approximately 300 Gb total), three SMRT Cells were used, with average Q-scores of Q33, Q33, and Q32, respectively.

COLO829 and COLO829BL sequencing data and COLO829BL DSA generation

The PacBio HiFi sequencing data of COLO829 (SMAFIJPL47BQ, SMAFIO69TOMB, SMAFIURDW1P6) and COLO829BL (SMAFIFRLAQ54) cell lines were generated as part of the SMaHT benchmarking study[51]. COLO829BL DSA was generated by UWSC GCC[24] using Verkko2[23]. All data were downloaded from the SMaHT data portal (https://data.smaht.org).

Pangenome graph construction and characterization

The assemblies from HPRC release 1[12], GRCh38, and CHM13 were constructed into v2 and v3 pangenome graphs using Minigraph-Cactus (v2.9.7)[15], with GRCh38 as reference. Nodes, edges, and edge degrees of three pangenome graphs were counted using Panacus (v0.2.4)[52]. The clipped and allele-filtered graphs were used for alignment and surjection. The clipped and unfiltered graphs were used for pangenome personalization. Minigraph-Cactus also generated the pangenome deconstructed VCF, which was used to visualize non-reference variants in Figure 3B.

Personalized pangenome (pangenome-inferred DSA) construction and evaluation

We used the HG005 simulated short-read WGS data and COLO829BL short-read WGS data to construct the personalized pangenome graphs from v3 pangenome graphs (clipped and not allele-filtered). We indexed the clipped graph using vg (v1.64.1) to create the index .hapl file, .ri file and .dist file:

```
vg index -t 10 -j $dist --no-nested-distance $gbz
vg gbwt -p --num-threads 10 -r $ri -Z $gbz
vg haplotypes -v 2 -t 10 -H $hapl $gbz
```

and utilized the KMC (v3.2.4)[53] with -k29 -m128 -okff to count the k-mers from the sample's WGS library. vg haplotype[17] was used for haplotype sampling and constructing the personalized pangenome graph.

```
vg haplotypes -v $level -t $threads --include-reference --diploid-sampling -i $hapl -k $kff -g
$output_gbz $gbz
```

This imputes a subgraph from the pangenome graphs with GRCh38 as reference and two recombinant assemblies (hap1, hap2). Two recombinant assemblies were extracted from graphs to FASTA format and used as pangenome-inferred assemblies.

The pangenome-inferred assemblies are evaluated using linear and graph-based approaches. The linear approach compares the number of base substitutions, insertions, and deletions between pangenome-inferred assemblies with HG005 maternal/paternal assemblies and that between GRCh38 and HG005 maternal/paternal assemblies. The assemblies were aligned with Minimap2 (v2.28)[20] using -cx asm5; --secondary=no settings to generate PAF files. The paftools.js stat was used to count the number of substitutions, insertions, and deletions from PAF files. The graph-based approach incorporated pangenome-inferred assemblies and the HG005 or COLO829BL diploid assemblies[24] into a five-assembly Minigraph-Cactus[15] validation graph, with GRCh38 as reference. The graphs were deconstructed to VCF with genotypes from each assembly, which were then used to compare genotypes between the pangenome-inferred assemblies and the HG005 or COLO829BL diploid assemblies and calculate the capture rate.

Assembly scaffolding

RagTag (v2.1.0)[54] was used to scaffold the assemblies (HG005 maternal and paternal assemblies, pangenome-inferred assemblies of HG005 and COLO829BL), using GRCh38 (GCA_000001405.29) as the reference assembly.

Whole-genome sequencing data processing for the HapMap mixture and COLO829

After raw sequence quality assessment, Illumina short reads were processed using fastp[55] to remove read pairs containing polyG artifacts. Illumina short reads of the HapMap mixture were aligned to linear references: GRCh38, HG005 paternal or maternal assemblies, and graph-based references: v1, v2, and v3 pangenome graphs (clipped and allele-filtered) and a personalized pangenome graph (subgraph contains GRCh38, pangenome-inferred assemblies of HG005 hap1 and hap2). GRCh38 (GCA_000001405.15) excludes all the ALT contigs and Human decoy sequences from version GCA_000786075.2 of GRCh38, including chromosomes from the GRCh38 primary assembly unit, mitochondrial genome from the GRCh38 non-nuclear assembly unit, unlocalized scaffolds from the GRCh38 primary assembly unit, and unplaced scaffolds from the GRCh38 primary assembly unit and Epstein-Barr virus (EBV) sequence. HG005 paternal and maternal assemblies were obtained from HPRC release 1[12]. chrX and chrM from HG005 maternal assembly were added to HG005 paternal assembly, and chrY from HG005 paternal assembly was added to HG005 maternal assembly to ensure the assemblies have a full set of chromosomes. The PacBio HiFi data of the HapMap mixture were aligned to linear references: GRCh38, HG005 paternal or maternal assemblies, and pangenome-inferred assemblies of HG005 hap1 or hap2. The PacBio HiFi data of COLO829 and COLO829BL were aligned to linear references: GRCh38, pangenome-inferred assemblies of COLO829BL hap1 or hap2.

Illumina short read data were aligned to the linear references using BWA-MEM (v0.7.17)[56]. PacBio HiFi data were aligned using pbmm2[20] (v1.13.1). The short-read data were aligned to the genome graphs using vg (v1.64.1) giraffe[16]. The reads were then surjected to GRCh38 (for v1, v2 and v3 pangenome graphs) or pangenome-inferred assemblies of HG005 (for personalized

pangenome graph) using vg surject[12]. For short-read data, the alignments were processed with Mark Duplicate, Indel Realignment, and Base Quality Score Recalibration (BQSR) using GATK (v4.1.0.0)[57] before variant calling.

Somatic variant calling and filtering for the HapMap mixture

SNVs and indels (small insertions and deletions less than 50 base pairs) were called from 500x short-read data using MuTect2 (v4.5.0.0)[18] tumor-only mode. The variants are then filtered using MuTect2 FilterMutectCalls[18]. The "haplotype" or "clustered_events" were accepted in addition to PASS FILTER tags to match the characteristics of the HapMap mixture, which has more than 10 million SNVs from 6 different cell lines and more than 6 million in the somatic VAF range, much more than expected in a healthy human sample. The SNVs and indels were decomposed and normalized by bcftools (v1.20-24-g5d48dbc3)[58] norm -a --atom-overlaps '.' --check-ref w -N and separated into SNVs and indels by Picard (v2.9.0) SplitVcfs.

SVs (insertions and deletions >= 50 base pairs) were called and filtered from 100x long-read data using Sniffles2 (v2.6.3) mosaic mode and set –mosaic-af-max 0 and –mosaic-af-max 0.218 following the developer's suggestion.[19] In Sniffles2 mosaic mode, the ALN_NM filter removes artifacts by excluding SVs whose reads show an unusually high length-weighted mismatch rate relative to the dataset average.[19] This filter was disabled because reduced genome-wide alignment mismatch rates in DSAs or pangenome-inferred DSAs (due to decreased reference bias) increase the filter stringency, making it incomparable with GRCh38.

Alignment metrics

The mapped reads, properly paired reads, gapless reads, and perfectly mapped reads were used as alignment metrics. Gapless reads are defined as alignments without insertions and deletions and might contain mismatches and/or softclips. Perfectly mapped reads are defined as alignments that map exactly to the reference. Properly paired reads are defined as alignments that paired-end reads must map, be in a faced orientation, and within a reasonable distance.

For alignments to linear references, the short-read alignment metrics were generated using samtools (v1.18)[58] stats, flagstats, and MD, CIGAR fields in BAM files. The long-read alignment metrics were generated using Qualimap (v2.2.1)[59]. For alignments to genome graphs, the short and long read alignment metrics were generated using vg stats from the GAM file.

Benchmarking variant sets generation

The benchmarking variant sets of the HapMap mixture based on GRCh38, pangenome-inferred assembly of HG005 (hap1 and hap2), and HG005 maternal and paternal assemblies were generated and validated using a genome-graph approach[14] (Figure S7). The assemblies of six HapMap samples were obtained from HPRC release 1[12]. The benchmarking variant sets contain all possible SNVs, indels, SVs from five somatic samples in the HapMap mixture, and benchmarking confident regions (assembly-reliable regions defined by HPRC).[14] The benchmarking confident regions for the pangenome-inferred assemblies of HG005 (hap1 and hap2) and the HG005 maternal and paternal assemblies were lifted over from the GRCh38 confident regions.

704 **Region difficulty stratification**
705
706 The GRCh38 stratification into easy, difficult, and extreme regions was generated by the SMaHT
707 network (https://data.smaht.org). The easy regions are based on 1KG strict mask[60]. The difficult
708 regions are regions outside of easy regions and within sample-agnostic easy regions defined using
709 a pan-genome approach[61]. The extreme regions are defined as those outside the easy and difficult
710 regions. The stratifications for pangenome-inferred DSA of HG005, HG005 maternal and paternal
711 assemblies were generated by lifting over regions from GRCh38 to each assembly using UCSC
712 liftOver with default settings. The CHAIN files were generated by the nf-LO pipeline (v1.8.6)[62]
713 with default settings. These regions are then overlapped with confident regions from benchmarking
714 variant sets using bedtools[63] -isec.
715
716 **Variant calling benchmark**
717 The benchmarking was performed within the benchmarking confident regions of chr1-22, chrX,
718 and chrY. SNVs and indels were benchmarked using RTG (v3.12.1) vcfeval[64] with –squash-ploidy
719 and –all-records. SVs were benchmarked using Truvari (v5.2.0)[65] with --pctsize 0.9, --pctseq 0.9,
720 --pick multi.
721
722 The somatic SNVs, indels, and SVs were partitioned into 9 VAF bins: <0.5%, 0.5-1%, 1-1.5%,
723 1.5-2%, 2-3%, 3-4%, 4-5%, 5-10%, 10-16.5%. For recall, the benchmarking variant set was split
724 by VAF, and each bin was compared with the entire call set, while for precision, the call set was
725 split by VAF and compared with the entire benchmarking variant set. The precision and recall were
726 calculated as: Precision = TP/(TP+FN); Recall = TP/(TP+FP).
727
728 **Somatic variant calling and mutational signature analysis for COLO829**
729 Somatic SNVs were identified using DeepSomatic (v1.8.0)[22] tumor–normal mode from COLO829
730 and COLO829BL PacBio HiFi data. The somatic SNV calling was performed using GRCh38,
731 pangenome-inferred DSA hap1 or hap2 of COLO829BL as the reference. The SNV calls with more
732 than one alternative allele supporting reads in COLO829BL BAM were removed in the initial call
733 sets. Next, the SNV calls from hap2 were lifted over to hap1 and compared with the hap1 call set
734 to remove the duplicated SNV calls. This harmonized call set was lifted over to GRCh38 and
735 compared with GRCh38 SNV calls. The lifted over SNVs that did not overlap with GRCh38 calls
736 and SNVs that failed to lift over were defined as pangenome-inferred DSA unique SNV calls.
737 Mutational signature analysis was performed on unique and shared SNV calls from GRCh38 or
738 pangenome-inferred DSAs with SigProfilerMatrixGenerator (v1.3.3)[66] and SigProfilerAssignment
739 (0.2.3)[67] to obtain SBS and DBS profiles and signatures based on COSMIC (v3.4)[68]
740 (https://github.com/ryansohny/VCF2SPECTRUM).
741
742 **Somatic SNV annotation and pathogenicity prediction for COLO829**
743 The somatic SNVs on GRCh38 were annotated with VEP (v115.2)[70]. The AlphaMissense[30]
744 pathogenicity scores were generated using VEP plugins. To annotate SNVs that cannot be lifted
745 over from pangenome-inferred DSA to GRCh38, the GRCh38 GENCODE annotation (v49)[71] was
746 mapped to pangenome-inferred DSA with LiftOff[72] with the following command:
747
748 liftoff -p 16 -g gencode.v49.primary_assembly.basic.annotation.gtf -o hs_hprc-v1.1-mc-
749 chm13_20x_1_gencode.v49.primary_assembly.basic.annotation_copies.gtf -u

750 unmapped_features.hs_hprc-v1.1-mc-chm13_20x_1_.txt -cds -copies hs_hprc-v1.1-mc-
751 chm13_20x_1.fasta ../GRCh38.primary_assembly.genome.fa
752
753 The somatic SNVs were then annotated with BCFtools/csq[73] to obtain variant consequences
754 (missense/synonymous). The k100 mappability for GRCh38 and pangenome-inferred DSA was
755 generated by GenMap[74] with the following commands:
756
757 genmap index -F hs_hprc-v1.1-mc-chm13_20x_1.fasta -I hs_hprc-v1.1-mc-
758 chm13_20x_1.genmap
759 genmap map -K 100 -E 0 -I hs_hprc-v1.1-mc-chm13_20x_1.genmap -O genmap_out --wig
760
761 **SNV mutation rate analysis for COLO829**
762 Pangenome-inferred DSA of COLO829 was annotated using RepeatMasker (v4.1.9)[69] using
763 default settings. The RepeatMasker annotation of GRCh38 was obtained from
764 https://www.repeatmasker.org/species/hg.html. The somatic SNVs in repetitive elements were
765 counted based on RepeatMasker annotations. The somatic SNVs in the exon or intron were
766 counted based on VEP annotations. The SNV mutation rate was calculated by dividing the total
767 SNV counts by the total region length.
768
769 **Data and code availability**
770 The whole genome sequencing data of HapMap mixture, COLO829, and COLO829BL, and
771 COLO829BL DSA generated by the SMaHT network can be found at https://data.smaht.org/.
772 The HapMap mixture benchmarking variant sets based on GRCh38, pangenome-inferred
773 assembly of HG005 (hap1 and hap2), and HG005 maternal and paternal assemblies can be found
774 in the repository on Zenodo: https://zenodo.org/records/17544617. All scripts are available at:
775 https://github.com/twlab/SMaHT_Pangenome_Benchmark.
776
777 **Acknowledgements**
778 We thank Genome Technology Access Center at McDonnell Genome Institute at Washington
779 University School of Medicine for sequencing data generation and management. We also thank all
780 members in the Somatic Mosaicism across Human Tissues Network. These studies were supported
781 through NIH grants UM1DA058219, 3UM1DA058219-01S1, U24NS132103, UM1HG010971,
782 and U41HG010972.
783
784 **Author contributions**
785 Q.F., Z.X., J.F.M., and T.W. contributed to study design and conceptualization. R.F., B.K.J.,
786 M.F.M., T.P., and H.J.J. performed cell culture and sequencing data production. T.L. managed
787 sequencing data producing pipeline. C.T. and D.L. managed sequencing data transfer. Q.F., Z.X.,
788 W.Z., N.K., Z.T., A.R., J.F.M., B.M., D.A., E.A.B., J.E.G., E.M., J.S., X.Z., S.D., Y.C., Z.L.
789 performed bioinformatics analyses. Q.F. and B.M. generated figures. Q.F., J.F.M., and Z.X. wrote
790 the initial manuscript. T.W., J.F.M., B.M., S.C.J., N.K., and Z.T. reviewed and edited the
791 manuscript. T.W., R.F., and H.S. administered the project. T.W., S.C.J., B.F., H.S., and H.A.L.
792 provided funding and resources.
793
794 **Declaration of interests**
795 The authors declare no competing interests.

References

1. Stratton, M.R., Campbell, P.J., and Futreal, P.A. (2009). The cancer genome. Nature *458*, 719–724.

2. Coorens, T.H.H., Oh, J.W., Choi, Y.A., and others (2025). The Somatic Mosaicism across Human Tissues Network. Nature *643*, 47–59.

3. Saini, N., and Gordenin, D.A. (2018). Somatic mutation load and spectra: A record of DNA damage and repair in healthy human cells. Environ Mol Mutagen *59*, 672–686.

4. Watson, I.R., Takahashi, K., Futreal, P.A., and Chin, L. (2013). Emerging patterns of somatic mutations in cancer. Nat Rev Genet *14*, 703–718.

5. Greenman, C., Stephens, P., Smith, R., and others (2007). Patterns of somatic mutation in human cancer genomes. Nature *446*, 153–158.

6. Olafsson, S., and Anderson, C.A. (2021). Somatic mutations provide important and unique insights into the biology of complex diseases. Trends in Genetics *37*, 872–881. https://doi.org/10.1016/j.tig.2021.06.012.

7. Sherman, R.M., and Salzberg, S.L. (2020). Pan-genomics in the human genome era. Nat Rev Genet *21*, 243–254. https://doi.org/10.1038/s41576-020-0210-7.

8. Lander, E.S., Linton, L.M., Birren, B., and others (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

9. Olson, N.D., Wagner, J., Dwarshuis, N., Miga, K.H., Sedlazeck, F.J., Salit, M., and Zook, J.M. (2023). Variant calling and benchmarking in an era of complete human genome sequences. Nat Rev Genet *24*, 464–483. https://doi.org/10.1038/s41576-023-00590-0.

10. Pollard, R.D., Wilkerson, M.D., and Rajagopal, P.S. (2024). Identification of germline population variants misclassified as cancer-associated somatic variants. Front. Med. *11*, 1361317. https://doi.org/10.3389/fmed.2024.1361317.

11. 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., et al. (2022). The Human Pangenome Project: a global resource to map genomic diversity. Nature *604*, 437–446. https://doi.org/10.1038/s41586-022-04601-8.

12. Liao, W.W., Asri, M., Ebler, J., and others (2023). A draft human pangenome reference. Nature *617*, 312–324.

13. Consortium, I.H. (2003). The International HapMap Project. Nature *426*, 789–796.

14. Kong, N., Tang, Z., Ruttenberg, A., Macias-Velasco, J.F., Li, Z., Zhang, W., Miao, B., Xin, Z., Fu, Q., Park, H., et al. (2025). A Pangenomic Method for Establishing a Somatic Variant

Detection Resource in HapMap Mixtures. bioRxiv, 2025.09.29.679336. https://doi.org/10.1101/2025.09.29.679336.

15. Hickey, G., Monlong, J., Ebler, J., Novak, A.M., Eizenga, J.M., Gao, Y., Human Pangenome Reference Consortium, Abel, H.J., Antonacci-Fulton, L.L., Asri, M., et al. (2024). Pangenome graph construction from genome alignments with Minigraph-Cactus. Nat Biotechnol *42*, 663–673. https://doi.org/10.1038/s41587-023-01793-w.

16. Sirén, J., Monlong, J., Chang, X., Novak, A.M., Eizenga, J.M., Markello, C., Sibbesen, J.A., Hickey, G., Chang, P.-C., Carroll, A., et al. (2021). Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. Science *374*, abg8871. https://doi.org/10.1126/science.abg8871.

17. Sirén, J., Eskandar, P., Ungaro, M.T., and others (2024). Personalized pangenome references. Nat Methods *21*, 2017–2023.

18. Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol *31*, 213–219. https://doi.org/10.1038/nbt.2514.

19. Smolka, M., Paulin, L.F., Grochowski, C.M., Horner, D.W., Mahmoud, M., Behera, S., Kalef-Ezra, E., Gandhi, M., Hong, K., Pehlivan, D., et al. (2024). Detection of mosaic and population-level structural variants with Sniffles2. Nat Biotechnol *42*, 1571–1580. https://doi.org/10.1038/s41587-023-02024-y.

20. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

21. Dou, Y., Gold, H.D., Luquette, L.J., and Park, P.J. (2018). Detecting Somatic Mutations in Normal Cells. Trends in Genetics *34*, 545–557. https://doi.org/10.1016/j.tig.2018.04.003.

22. Park, J., Cook, D.E., Chang, P.-C., Kolesnikov, A., Brambrink, L., Mier, J.C., Gardner, J., McNulty, B., Sacco, S., Keskus, A.G., et al. (2025). Accurate somatic small variant discovery for multiple sequencing technologies with DeepSomatic. Nat Biotechnol. https://doi.org/10.1038/s41587-025-02839-x.

23. Antipov, D., Rautiainen, M., Nurk, S., Walenz, B.P., Solar, S.J., Phillippy, A.M., and Koren, S. (2025). Verkko2 integrates proximity-ligation data with long-read De Bruijn graphs for efficient telomere-to-telomere genome assembly, phasing, and scaffolding. Genome Res. *35*, 1583–1594. https://doi.org/10.1101/gr.280383.124.

24. Sohn, M.-H., Dubocanin, D., Vollger, M.R., Kwon, Y., Minkina, A., Munson, K.M., Hart, S.F., Ranchalis, J.E., Parmalee, N.L., Sedeno-Cortes, A.E., et al. (2025). A telomere-to-telomere map of somatic mutation burden and functional impact in cancer. bioRxiv, 2025.10.10.681725. https://doi.org/10.1101/2025.10.10.681725.

867 25. Barra, V., and Fachinetti, D. (2018). The dark side of centromeres: types, causes and
868 consequences of structural abnormalities implicating centromeric DNA. Nat Commun *9*,
869 4340. https://doi.org/10.1038/s41467-018-06545-y.

870 26. Scelfo, A., Angrisani, A., Grillo, M., Barnes, B.M., Muyas, F., Sauer, C.M., Leung, C.W.B.,
871 Dumont, M., Grison, M., Mazaud, D., et al. (2024). Specialized replication mechanisms
872 maintain genome stability at human centromeres. Molecular Cell *84*, 1003-1020.e10.
873 https://doi.org/10.1016/j.molcel.2024.01.018.

874 27. Chen, S., Francioli, L.C., Goodrich, J.K., Collins, R.L., Kanai, M., Wang, Q., Alföldi, J.,
875 Watts, N.A., Vittal, C., Gauthier, L.D., et al. (2024). A genomic mutational constraint map
876 using variation in 76,156 human genomes. Nature *625*, 92–100.
877 https://doi.org/10.1038/s41586-023-06045-0.

878 28. Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G.,
879 Garcia, F., Young, N., et al. (2013). The Genotype-Tissue Expression (GTEx) project. Nat
880 Genet *45*, 580–585. https://doi.org/10.1038/ng.2653.

881 29. The Cancer Genome Atlas Research Network, Weinstein, J.N., Collisson, E.A., Mills, G.B.,
882 Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M.
883 (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet *45*, 1113–1120.
884 https://doi.org/10.1038/ng.2764.

885 30. Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulytė, A., Applebaum, T., Pritzel, A., Wong,
886 L.H., Zielinski, M., Sargeant, T., et al. (2023). Accurate proteome-wide missense variant
887 effect prediction with AlphaMissense. Science *381*, eadg7492.
888 https://doi.org/10.1126/science.adg7492.

889 31. Juan-Guadarrama, D.G., Beltrán-Navarro, Y.M., Reyes-Cruz, G., and Vázquez-Prado, J.
890 (2023). Ephexin3/ARHGEF5 Together with Cell Migration Signaling Partners within the
891 Tumor Microenvironment Define Prognostic Transcriptional Signatures in Multiple Cancer
892 Types. IJMS *24*, 16427. https://doi.org/10.3390/ijms242216427.

893 32. Patkunarajah, A., Stear, J.H., Moroni, M., Schroeter, L., Blaszkiewicz, J., Tearle, J.L., Cox,
894 C.D., Fürst, C., Sánchez-Carranza, O., Ocaña Fernández, M.D.Á., et al. (2020).
895 TMEM87a/Elkin1, a component of a novel mechanoelectrical transduction pathway,
896 modulates melanoma adhesion and migration. eLife *9*, e53308.
897 https://doi.org/10.7554/eLife.53308.

898 33. Bao, J., Wang, S., Gunther, L.K., Kitajiri, S., Li, C., and Sakamoto, T. (2015). The actin-
899 bundling protein TRIOBP-4 and -5 promotes the motility of pancreatic cancer cells. Cancer
900 Letters *356*, 367–373. https://doi.org/10.1016/j.canlet.2014.08.005.

901 34. Shen, T., Zhang, J., Wang, Y., and Liu, B. (2022). FCGBP Is a Promising Prognostic
902 Biomarker and Correlates with Immunotherapy Efficacy in Oral Squamous Cell Carcinoma.
903 Journal of Immunology Research *2022*, 1–11. https://doi.org/10.1155/2022/8443392.

35. Andries, V., Vandepoele, K., Staes, K., Berx, G., Bogaert, P., Van Isterdael, G., Ginneberge, D., Parthoens, E., Vandenbussche, J., Gevaert, K., et al. (2015). NBPF1, a tumor suppressor candidate in neuroblastoma, exerts growth inhibitory effects by inducing a G1 cell cycle arrest. BMC Cancer *15*, 391. https://doi.org/10.1186/s12885-015-1408-5.

36. Mose, L.E., Perou, C.M., and Parker, J.S. (2019). Improved indel detection in DNA and RNA via realignment with ABRA2. Bioinformatics *35*, 2966–2973. https://doi.org/10.1093/bioinformatics/btz033.

37. Cárdenas-Navia, L.I., Cruz, P., Lin, J.C., Nisc Comparative Sequencing Program, Rosenberg, S.A., and Samuels, Y. (2010). Novel somatic mutations in heterotrimeric G proteins in melanoma. Cancer Biology & Therapy *10*, 33–37. https://doi.org/10.4161/cbt.10.1.11949.

38. Mi, J., Han, Y., Zhang, J., Hao, X., Xing, M., and Shang, C. (2020). Long noncoding RNA LINC01410 promotes the tumorigenesis of neuroblastoma cells by sponging microRNA-506-3p and modulating WEE1. Cancer Medicine *9*, 8133–8143. https://doi.org/10.1002/cam4.3398.

39. Du, Q., Zhang, M., Gao, A., He, T., and Guo, M. (2024). Epigenetic silencing *ZSCAN23* promotes pancreatic cancer growth by activating Wnt signaling. Cancer Biology & Therapy *25*, 2302924. https://doi.org/10.1080/15384047.2024.2302924.

40. Kim, J., Hwang, H.-G., Jeon, H.-J., Kim, S.I., Kim, M., and Kim, J.-Y. (2024). ARHGEF5 binds Drebrin and affects α-tubulin acetylation to direct neuronal morphogenesis and migration during mouse brain development. Front. Mol. Neurosci. *17*, 1421932. https://doi.org/10.3389/fnmol.2024.1421932.

41. Komiya, Y., Onodera, Y., Kuroiwa, M., Nomimura, S., Kubo, Y., Nam, J.-M., Kajiwara, K., Nada, S., Oneyama, C., Sabe, H., et al. (2016). The Rho guanine nucleotide exchange factor ARHGEF5 promotes tumor malignancy via epithelial–mesenchymal transition. Oncogenesis *5*, e258–e258. https://doi.org/10.1038/oncsis.2016.59.

42. Xu, H., Chen, D., Lu, J., and Zhong, L. (2024). High expression of ARHGEF5 predicts unfavorable prognosis in acute myeloid leukemia. Discov Onc *15*, 491. https://doi.org/10.1007/s12672-024-01364-8.

43. Alliance of Genome Resources https://www.alliancegenome.org.

44. Mo, S., Zhang, L., Dai, W., Han, L., Wang, R., Xiang, W., Wang, Z., Li, Q., Yu, J., Yuan, J., et al. (2020). Antisense lncRNA LDLRAD4-AS1 promotes metastasis by decreasing the expression of LDLRAD4 and predicts a poor prognosis in colorectal cancer. Cell Death Dis *11*, 155. https://doi.org/10.1038/s41419-020-2338-y.

45. Qin, Z., Ren, F., Xu, X., Ren, Y., Li, H., Wang, Y., Zhai, Y., and Chang, Z. (2009). ZNF536, a Novel Zinc Finger Protein Specifically Expressed in the Brain, Negatively Regulates Neuron Differentiation by Repressing Retinoic Acid-Induced Gene Transcription. Molecular and Cellular Biology *29*, 3633–3643. https://doi.org/10.1128/MCB.00362-09.

46. Fang, Y., Dong, R., Wang, S., Chen, L., Wang, Y., and Wu, Q. (2025). Zinc Finger Protein 536 is a Potential Prognostic Biomarker that Promotes Neuroblastoma Progression via VEGFR2-PI3K-AKT Pathway. CCDT *25*, 1447–1460. https://doi.org/10.2174/0115680096332753240913104409.

47. Meyskens, F.L., Edwards, L., and Levine, N.S. (1986). Role of topical tretinoin in melanoma and dysplastic nevi. Journal of the American Academy of Dermatology *15*, 822–825. https://doi.org/10.1016/S0190-9622(86)70239-9.

48. Li, L., Chen, S., Tang, Y., Wu, J., He, Y., and Qiu, L. (2022). Oncogene or tumor suppressor gene: An integrated pan-cancer analysis of NBPF1. Front. Endocrinol. *13*, 950326. https://doi.org/10.3389/fendo.2022.950326.

49. Park, S., Lee, H., Kim, M., Park, J., Kim, S.-H., and Park, J. (2018). Emerging roles of TRIO and F-actin-binding protein in human diseases. Cell Commun Signal *16*, 29. https://doi.org/10.1186/s12964-018-0237-y.

50. Liu, Q., Niu, X., Li, Y., Zhang, J., Zhu, S., Yang, Q., Zhang, W., and Gong, L. (2022). Role of the mucin-like glycoprotein FCGBP in mucosal immunity and cancer. Front. Immunol. *13*, 863317. https://doi.org/10.3389/fimmu.2022.863317.

51. The Somatic Mosaicism across Human Tissues Network, and Abyzov, A. (2025). Comprehensive benchmarking of somatic mutation detection by the SMaHT Network. bioRxiv, 2025.10.09.678885. https://doi.org/10.1101/2025.10.09.678885.

52. Parmigiani, L., Garrison, E., Stoye, J., Marschall, T., and Doerr, D. (2024). Panacus: fast and exact pangenome growth and core size estimation. Bioinformatics *40*, btae720. https://doi.org/10.1093/bioinformatics/btae720.

53. Kokot, M., Długosz, M., and Deorowicz, S. (2017). KMC 3: counting and manipulating *k* -mer statistics. Bioinformatics *33*, 2759–2761. https://doi.org/10.1093/bioinformatics/btx304.

54. Alonge, M., Lebeigle, L., Kirsche, M., Jenike, K., Ou, S., Aganezov, S., Wang, X., Lippman, Z.B., Schatz, M.C., and Soyk, S. (2022). Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. Genome Biol *23*, 258. https://doi.org/10.1186/s13059-022-02823-7.

55. Chen, S. (2025). fastp 1.0: An ultra-fast all-round tool for FASTQ data quality control and preprocessing. iMeta, e70078. https://doi.org/10.1002/imt2.70078.

56. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

57. Van Der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. CP in Bioinformatics *43*. https://doi.org/10.1002/0471250953.bi1110s43.

58. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., et al. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*, giab008. https://doi.org/10.1093/gigascience/giab008.

59. Okonechnikov, K., Conesa, A., and García-Alcalde, F. (2016). Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics *32*, 292–294. https://doi.org/10.1093/bioinformatics/btv566.

60. The 1000 Genomes Project Consortium, Corresponding authors, Auton, A., Abecasis, G.R., Steering committee, Altshuler, D.M., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74. https://doi.org/10.1038/nature15393.

61. Li, H. (2025). Finding easy regions for short-read variant calling from pangenome data. GigaScience *14*, giaf103. https://doi.org/10.1093/gigascience/giaf103.

62. Talenti, A., and Prendergast, J. (2021). nf-LO: A Scalable, Containerized Workflow for Genome-to-Genome Lift Over. Genome Biology and Evolution *13*, evab183. https://doi.org/10.1093/gbe/evab183.

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

64. Cleary, J.G., Braithwaite, R., Gaastra, K., Hilbush, B.S., Inglis, S., Irvine, S.A., Jackson, A., Littin, R., Rathod, M., Ware, D., et al. (2015). Comparing Variant Call Files for Performance Benchmarking of Next-Generation Sequencing Variant Calling Pipelines. Preprint at Bioinformatics, https://doi.org/10.1101/023754 https://doi.org/10.1101/023754.

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

66. Bergstrom, E.N., Huang, M.N., Mahto, U., Barnes, M., Stratton, M.R., Rozen, S.G., and Alexandrov, L.B. (2019). SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events. BMC Genomics *20*, 685. https://doi.org/10.1186/s12864-019-6041-2.

67. Díaz-Gay, M., Vangara, R., Barnes, M., Wang, X., Islam, S.M.A., Vermes, I., Duke, S., Narasimman, N.B., Yang, T., Jiang, Z., et al. (2023). Assigning mutational signatures to individual samples and individual somatic mutations with SigProfilerAssignment. Bioinformatics *39*, btad756. https://doi.org/10.1093/bioinformatics/btad756.

68. Tate, J.G., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M., Bindal, N., Boutselakis, H., Cole, C.G., Creatore, C., Dawson, E., et al. (2019). COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Research *47*, D941–D947. https://doi.org/10.1093/nar/gky1015.

1014 69. Smit, A.F., Hubley, R., and Green, P. (2013). RepeatMasker.

1015 70. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and
1016 Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol *17*, 122.
1017 https://doi.org/10.1186/s13059-016-0974-4.

1018 71. Mudge, J.M., Carbonell-Sala, S., Diekhans, M., Martinez, J.G., Hunt, T., Jungreis, I.,
1019 Loveland, J.E., Arnan, C., Barnes, I., Bennett, R., et al. (2025). GENCODE 2025: reference
1020 gene annotation for human and mouse. Nucleic Acids Research *53*, D966–D975.
1021 https://doi.org/10.1093/nar/gkae1078.

1022 72. Shumate, A., and Salzberg, S.L. (2021). Liftoff: accurate mapping of gene annotations.
1023 Bioinformatics *37*, 1639–1643. https://doi.org/10.1093/bioinformatics/btaa1016.

1024 73. Danecek, P., and McCarthy, S.A. (2017). BCFtools/csq: haplotype-aware variant
1025 consequences. Bioinformatics *33*, 2037–2039. https://doi.org/10.1093/bioinformatics/btx100.

1026 74. Pockrandt, C., Alzamel, M., Iliopoulos, C.S., and Reinert, K. (2020). GenMap: ultra-fast
1027 computation of genome mappability. Bioinformatics *36*, 3687–3692.
1028 https://doi.org/10.1093/bioinformatics/btaa222.

1029