Minimizing Reference Bias with an Impute-First Approach

Naga Sai Kavya Vaddadi[1,*], Taher Mun[1], and Ben Langmead[1,*]

[1]Department of Computer Science, Johns Hopkins University
[*]*corresponding authors;* kvaddad1@jhu.edu, langmea@cs.jhu.edu

December 1, 2023

Abstract

Pangenome indexes reduce reference bias in sequencing data analysis. However, a greater reduction in bias can be achieved using a personalized reference, e.g. a diploid human reference constructed to match a donor individual's alleles. We present a novel impute-first alignment framework that combines elements of genotype imputation and pangenome alignment. It begins by genotyping the individual from a subsample of the input reads. It next uses a reference panel and efficient imputation algorithm to impute a personalized diploid reference. Finally, it indexes the personalized reference and applies a read aligner, which could be a linear or graph aligner, to align the full read set to the personalized reference. This framework has higher variant-calling recall (99.54% vs. 99.37%), precision (99.36% vs. 99.18%), and F1 (99.45% vs. 99.28%) compared to a graph-based pangenome. The personalized reference is also smaller and faster to query compared to a pangenome index, making it an overall advantageous choice for whole-genome DNA sequencing experiments.

1 Introduction

The field of Bioinformatics has developed various ways to reduce the "reference bias" that results when aligning reads to reference genomes. Reference bias occurs when reads containing non-reference alleles fail to align to their true point of origin, leading to inaccurate results for analyses concerned with hypervariable regions [1], allele-specific effects [2–5], ancient DNA [6, 7] or epigenomics [8]. Pangenome alignment methods reduce bias by using a pangenome graph [9–11] or collection of linear genomes [12]. Pangenome indexes allow the aligner to remove unwanted penalties associated with known genetic variants. While these methods reduce bias [13–15], personalized references, which are tailored to include the specific alleles present in the individual under study (the "donor") are known to be better at reducing bias [14, 15].

Here we present an efficient "impute-first alignment" framework combining advantages of genotype imputation and pangenome alignment. Modern imputation tools are efficient [16, 17] and leverage comprehensive genetic and linkage-disequilibrium information from reference panels like the 1000 Genomes Project (1KGP) [18], which spans 1000s of haplotypes, and Human Genome Structural Variation Consortium (HGSVC2) [19], which includes complex structural variations. Importantly, these panels are phased, meaning they faithfully represent the co-inheritance patterns of nearby variants.

Having imputed a diploid personalized reference, we then align reads to it with a pangenome aligner, where the "pangenome" contains exactly the alleles imputed for the donor. This reduces reference bias while also reducing computational overhead relative to methods that use a large pangenome.

Past methods used personalized references to alleviate reference bias in the study of RNA-seq [4, 20, 21], but rely on the user to provide the donor's genotypes as an input. Some methods additionally perform a personalization step, but have other drawbacks. Gramtools [22] builds a pangenome reference representation that can be used both to align reads and to perform imputation of a personalized genome. That study was limited to monoploids and showed that a personalized genome for Plasmodium falciparum helps to alleviate reference bias in hypervariable regions. iCORN [23] gradually refines the reference to contain more ALT alleles through an expensive iterative alignment process, and also focuses on monoploids. MMSeq

[24] performs variant calling on an input BAM file to construct a personalized diploid transcriptome, but does not perform imputation. RefEditor [25] uses a genotype imputation tool (MaCH or Minimac) to obtain phased diploid genotypes based on SNP genotyping array data. By contrast, our work performs personalization based on the input (sequencing) dataset itself and does not require a second dataset. Importantly, none of the above tools used an efficient imputation method like Glimpse [17], which has the positional Burrows Wheeler Transform (PBWT) [26] as its computational engine. Further, none of the above-mentioned studies assessed the computational cost, and thus the practicality, of an efficient personalization phase that runs as part of the main data analysis.

Groza et al. [8] introduced Graph Personal Genome, or a graph-shaped reference containing only the donor's diploid variants. In their case, use of a personalized reference is a first step toward obtaining less biased ChIP-seq peak calls. Our framework also uses this idea. However, our framework both imputes the diploid variants and creates this graph-genome reference as part of the same data analysis. The Groza et al. framework takes the variants as an input, with the expectation that these would be called from a separate dataset. Also unlike Groza et al. we compare the method of constructing a personalized graph-genome reference to the more common practice of simply using a pangenome reference built over a large panel, i.e. the 1000 Genomes Project phase-3 callset.

We show that our impute-first framework can construct a diploid personalized reference at a low computational cost relative to the indexing of a pangenome. Further, we show that these personalized references are accurate, having high variant-level precision and recall, and accurate overall phasing. Finally, we show that a pangenome index built over a personalized genome is more computationally efficient, less biased, and produces more accurate downstream variant calls compared to a pangenome index, making the impute-first framework an appealing choice for whole-genome DNA sequencing data analyses.

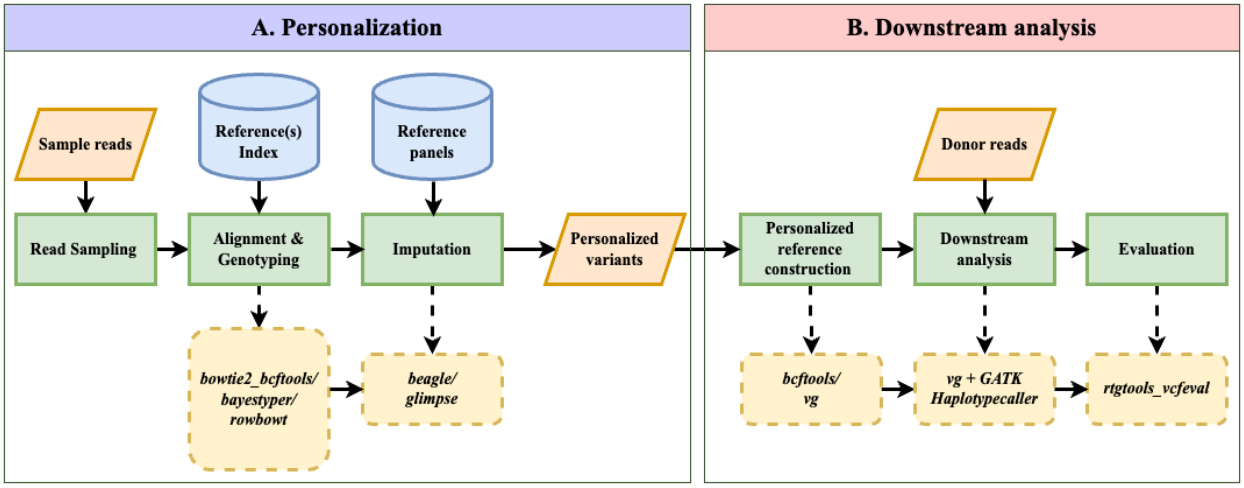


Figure 1: Impute-first alignment workflow in the context of analyzing a human whole-genome sequencing dataset. Box A shows the workflow up until it creates a personalized diploid reference. This the "personalization" component. Box B continues the workflow by aligning the full set of reads to the personalized reference. This is the "downstream" component.

2 Results

2.1 Overview of Impute-first Alignment Workflow

Figure 1 illustrates the workflow, which is divided into two components. In the first, a sample of the input reads is analyzed using a pangenome genotyper (e.g. Rowbowt [32]) and imputation tool (e.g. Glimpse [17]). The output is a phased personalized variant callset, which in turn is used to construct a personalized diploid reference genome. In the second component, the personalized reference is indexed (e.g. using VG

Personalization			
Step	Input	Tool (designator)	Output
Read Sampling	**Donor reads**: Whole-genome DNA-seq reads from HG001/NA12878, accession SRR622457 [18]	`seqtk` [27]	**Reads** sampled to 0.01, 0.05, 0.1, 0.2, 0.5, 1, 2, and 5× average coverage
Alignment & Genotyping	**Genotype panel**: HGSVC2 [19] VCF, excluding HG001/NA12878 and family members. **Reference**: GRCh38 primary assembly [28] **Reads**: Output from Sampling step	`bowtie2`[29] + `bcftools` (bb) [30]; `bayestyper`[31] (bt); `rowbowt`[32] (r);	**Rough genotype calls** in VCF format
Imputation	**Imputation panel & reference**: Same as previous step **Genotype calls**: Output from genotyping step.	`beagle`[16] (b); `glimpse`[17] (g)	**Personalized reference** as phased VCF file
Downstream analysis			
Step	Input	Tool	Output
Personalized Reference Construction	**Personalized reference**: from Imputation step **Reference**: GRCh38 primary assembly [28]	`bcftools`; `vg autoindex` [10]	**Personalized reference** as diploid FASTA or graph
Downstream Analysis	**Personalized reference**: From above step **Donor reads**: HG001/NA12878, acc. SRR622457 [18]	`vg giraffe` [10] `gatk HaplotypeCaller` [33]	**Variant calls** as VCF
Evaluation	**True variants**: HG001 VCF from GIAB [34]; High-confidence region annotations, etc.	`rtgtools` `vcfeval` [35]	**Benchmarking metrics**

Table 1: The steps, inputs, outputs and tools used in the tested Impute-first workflows.

Giraffe) and a read aligner analyzes the full read set with respect to this index. The workflow is modular; different tools can be substituted for the initial genotyping step (e.g. Bowtie2+bcftools instead of Rowbowt), the imputation step (e.g. Beagle instead of Glimpse) and the final read alignment step (e.g. Bowtie2 or BWA-MEM instead of VG Giraffe). The table uses underlining to highlight short "designators" that we use later to refer to particular tools, e.g. "r" for Rowbowt and "bb" for Bowtie2+bcftools.

2.2 Personalization

We studied personalizations workflows that used various algorithms and reference representations. The Bowtie2+bcftools workflow used a linear reference index, Rowbowt used a collection of linear haplotypes indexed with r-index, and BayesTyper used a graph pangenome index. We evaluated the workflows' abilities to efficiently call correct phased genotypes. We used a subsample of the reads from a real dataset containing PCR-free whole-genome DNA sequence reads for individual HG001/NA12878 (1000 Genomes Project [18] accession SRR622457). The initial sequencing depth was $30\times$, which we subsampled to lower coverage levels ranging from $0.01\times$ to $5\times$ (Table 1). The subsampled read sets help to measure the workflow's robustness to smaller inputs, which is relevant for scenarios where (a) the input dataset is small to begin with, or (b) we are seeking to minimize time and memory taken by the personalization steps by running on only a subsample of the full input. We hypothesized that we could still achieve high personalization accuracy on subsamples based on a previous study [36]. True NA12878/HG001 genotypes were taken from the HGSVC2 reference panel [19, 37]. We measured different personalization workflows' ability to correctly call alternate alleles (ALTs) and heterozygous variants (HETs). We also measured window accuracy, i.e. the fraction of 200 base-pair windows in the imputed diploid genome that exactly match the same window in the true genome.

Call Accuracy We assessed genotype accuracy across various coverages both before and after imputation. We measured precision, recall, and F1 score, i.e. the harmonic mean of precision and recall. As seen in Figure 2, increased coverage led to more accurate calls in all cases. Also, imputation led to notable improvements in F1 for both ALT and HET calls and for all three genotyping methods.

In most scenarios, Rowbowt produced more accurate genotypes and imputed genotypes compared to the other methods. An exception was HET variants as imputed by Beagle, where the Bowtie2+bcftools workflow achieved a somehwat higher F1. BayesTyper generally produced less accurate genotypes and

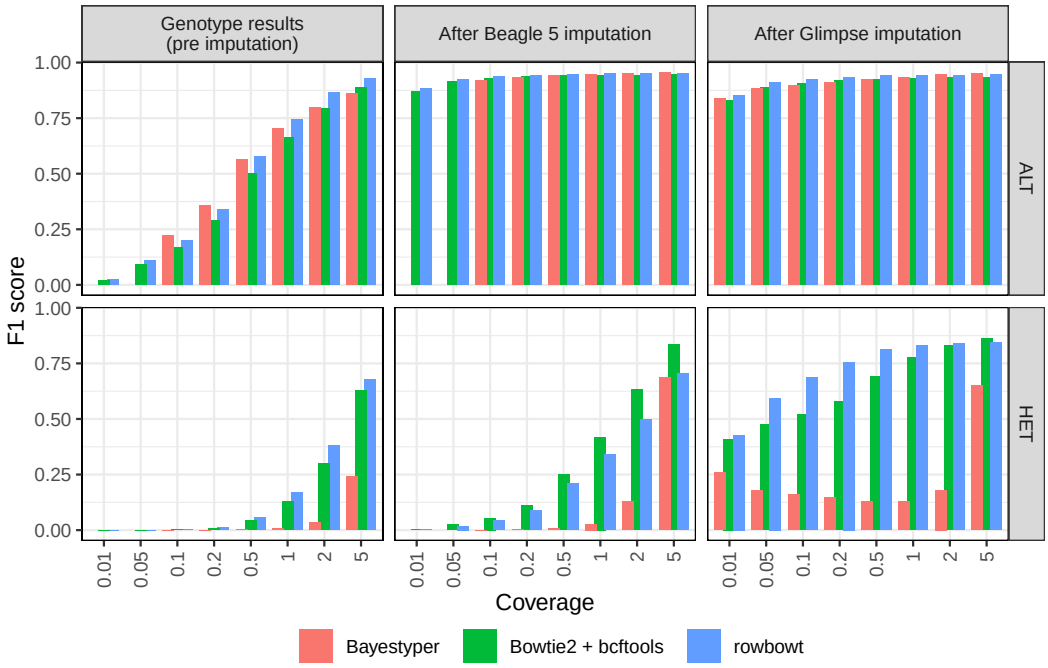


Figure 2: Aggregate F1 scores of alternate allele calls (ALT) and heterozygous calls (HET) across all variant types, generated using each alignment/genotyping method in the Impute-first alignment workflow.

imputed genotypes compared to the other tools, though this gap closed somewhat at higher coverages. For all three genotyping methods, Glimpse produced higher F1 for HET calls than Beagle.

Window Accuracy Window accuracy is a measure of how frequently a window of genomic bases covering a polymorphic site is correctly inferred in the personalized reference. We considered each of the 18.8M sites that were polymorphic according to the HGSVC2 reference panel. We call the site under consideration the "pivot." As a group, we consider the pivot as well as all other polymorphic sites to the left of and within 200 bp of the pivot. For each such group, we determined whether all variant calls for all variants in the group were correctly genotyped and correctly phased. Figure 3 illustrates the results, with windows stratified according to the total number of polymorphic sites falling inside. The "1 – 5" stratum includes groups with 1–5 polymorphic sites, "6 – 10" stratum includes groups with 6–10 polymorphic sites, and "11+" includes groups with 11 or more polymorphic sites. As expected, window accuracy increases with average coverage, since more total evidence leads to more accurate genotyping calls. Also as expected, window accuracy decreases as we consider denser strata (11+ being the densest), reflecting the increased difficulty of correctly calling all phased genotypes for all variants in a large group.

Rowbowt generally achieved the highest overall fraction of correctly-called windows. Rowbowt performed particularly well at $\leq$ 1x coverage. Conversely, in regions with $>$ 1x coverage, when used in combination with Beagle, Bowtie2+bcftools achieves higher (sometimes highest overall) accuracy.

Computational performance Figure 4 shows time and memory taken by these tool combinations. Glimpse used consistently less time and memory than Beagle. Specifically, Glimpse finished in about half the time used 1/100th as much memory as Beagle on average (Note the scale differences for the memory plots in Figure 4). Workflows using the Rowbowt genotyping tool were the fastest for the Genotyping step. For instance, at $1\times$ coverage, the Rowbowt genotyping step took 22 minutes, whereas Bowtie2+bcftools took 112 minutes and BayesTyper took 120 minutes. Workflows using Bowtie2+bcftools were the most memory-efficient. At $1\times$ coverage, Bowtie2+bcftools used 3.44 GB peak resident set size, while Rowbowt took 14.5 GB and BayesTyper took 30.2 GB.

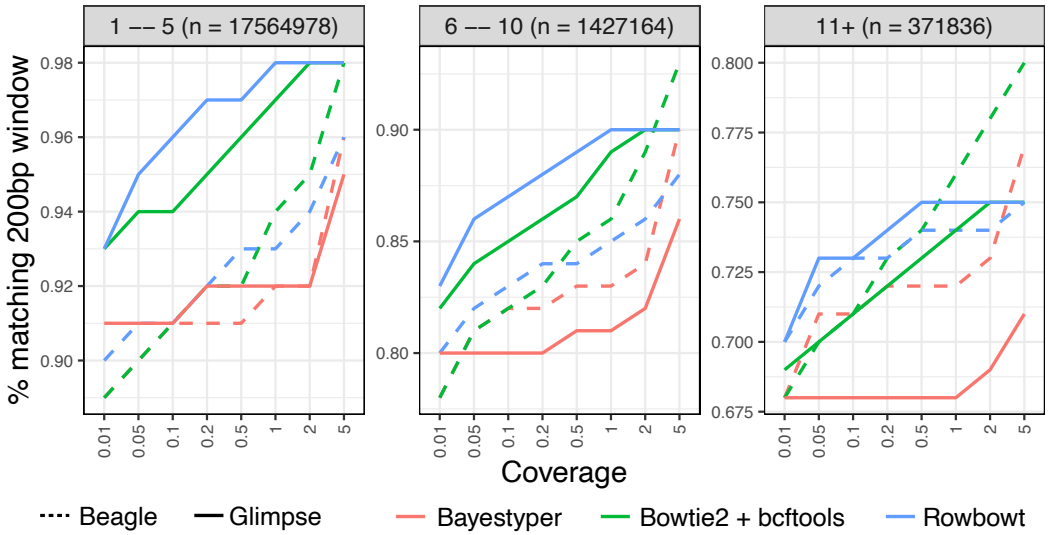


Figure 3: Window accuracy for diploid personalized genomes. The imputed sequence for each 200-bp windows anchored to a polymorphic site was compared to truth NA12878 sequence. Results are stratified by the number of polymorphic sites in the window.

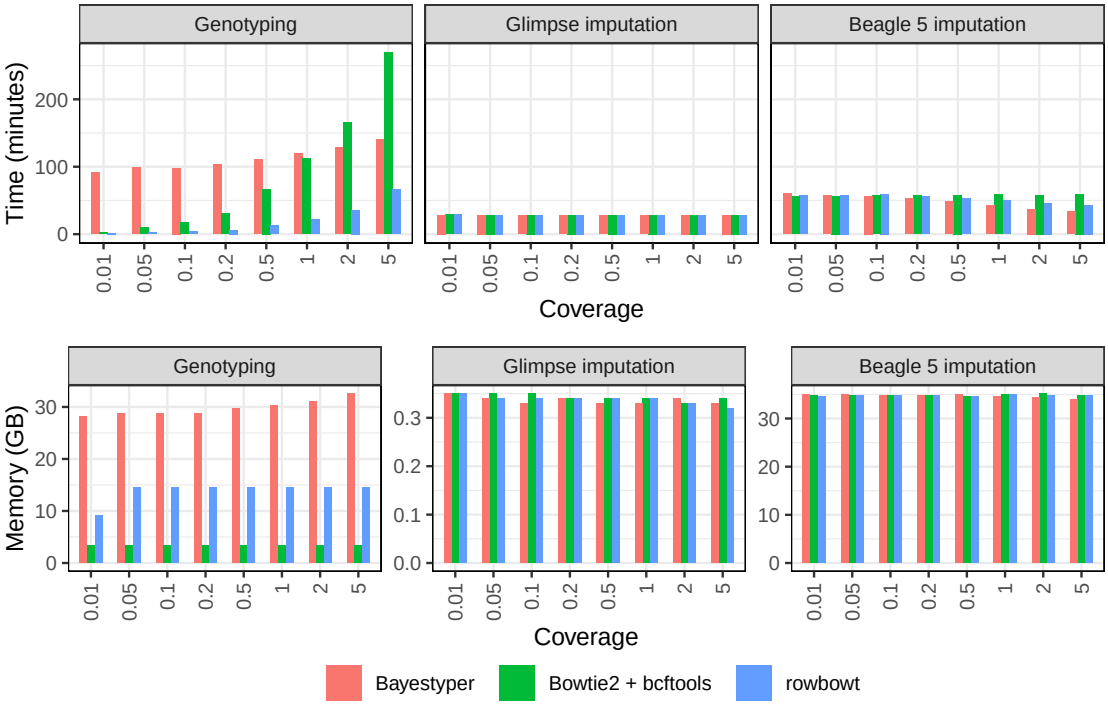


Figure 4: Computational overhead of genotyping (left) and imputation (middle: Glimpse, right: Beagle) in the personalization component of the impute-first workflow. In all cases, three genotyping methods are benchmarked. The mputation methods are benchmarked three times, using the inputs from each of the 3 genotyping methods. Note the differences in scale for the plots in the bottom row.

Choices for Downstream Analysis Given these results, we chose a few tool combinations for further analysis. Specifically, we chose the combinations called Rowbowt-Glimpse-1x (rgc1), Rowbowt-Glimpse-5x (rgc5), Bowtie2+bcftools-Glimpse-5x (bbgc5), and Bowtie2+bcftools-Beagle-5x (bbbc5). We chose these based on a combination of their high window accuracy, with $1\times$ and $5\times$ being relatively sparser subsamples that nonetheless achieved high calling accuracy. We refer to these tool combinations by their short designators: rgc1, rgc5, bbbc5, or bbgc5.

2.3 Impact of personalized references on downstream results

Having produced personalized diploid references with the rgc1, rgc5, bbbc5, and bbgc5 personalization workflows, we next evaluated downstream results obtained using the personalized references. We compared the performance of downstream workflows using both personalized references as well as typical linear and graph-pangenome references. We used a combination of the BWA-MEM and VG Giraffe aligners, as seen in Table 2. We used a sequencing dataset from individual HG001 consisting of 361 million 151 bp paired-end real reads (Illumina NovaSeq PCR-free, SRR14724533) and totaling $30\times$ average coverage.

Aligner	Reference type	Used for
BWA-MEM	rgc1-imputed diploid	Alignment score
BWA-MEM	Linear reference	Alignment score & variant calls
VG Giraffe	Linear reference	Allelic balance & variant calls
	1000 Genomes phase-3 pangenome graph*	
	rgc1-imputed diploid	
	rgc5-imputed diploid	
	bbgc5-imputed diploid	
	bbbc5-imputed diploid	

Table 2: "Downstream workflows" assessed. * The 1000 Genomes phase-3 pangenome excluded NA12878, as well as family members: NA12889, NA12890.

2.3.1 Alignment scores

We compared two sets of read alignments: one aligned to the the personalized diploid reference produced by the rgc1workflow, and another aligned to the linear GRCh38 reference ("GRC" for short). For each read that aligned to both, we asked whether its alignment score differed when aligned to the GRC versus to the personalized reference. 7.2% of the input reads had a different alignment score when aligned to one reference versus the other. Of these, 97% had higher alignment scores when aligned to the personalized reference. Among these reads, the alignment score improved by 7.7 points on average, equivalent in BWA-MEM's scoring scheme to about 2 mismatches (each with a penalty of 4) or 1 small indel (with an open penalty of 6 and extension penalty of 1). Figure 5 shows all the alignment-score differences for reads having non-zero difference. These experiments show that by aligning to the personalized reference, we frequently remove penalties associated with the presence of non-reference alleles in the donor reads.

2.3.2 Allelic balance at HETs

Next, we sought to measure how the personalized reference improves our ability to achieve a balanced representation of REF and ALT alleles at heterozygous (HET) sites in the donor. To do this, we began by using the Genome-in-a-Bottle (GIAB) HG001 v4.2 truth VCF file to identify sites where the NA12878/HG001 individual has a heterozygous varient. Following this identification, we used Biastools [38] to measure the

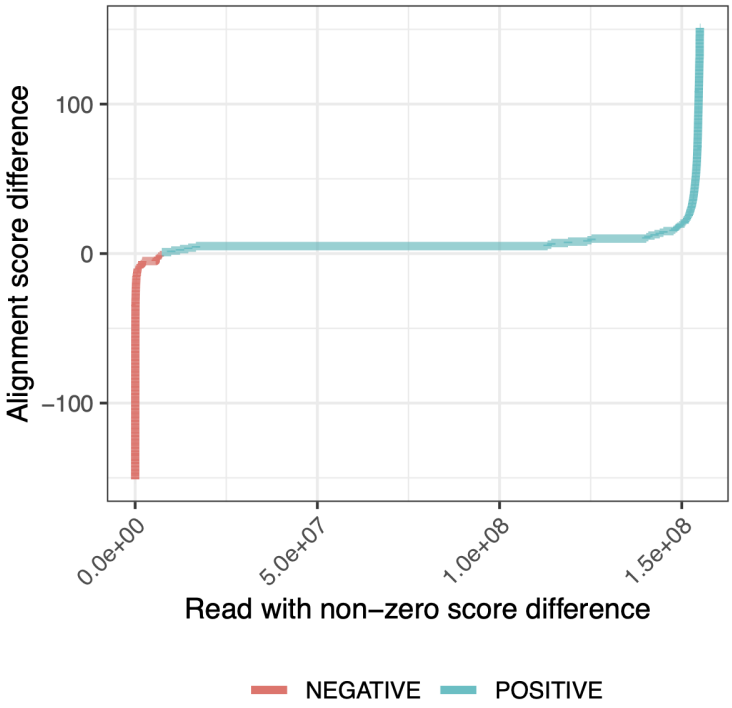


Figure 5: Difference in HG001 donor read alignment scores between personalized diploid reference ("Impute-first: rgc1")and linear reference genome (GRCh38).

allelic balance at each of these sites. Subsequently, we created an indel-balance plot, similar to those seen in previous studies of reference bias. Notably, we subsetted the analysis to focus on alignment intervals corresponding to the GIAB HG001 High-Confidence Regions (Figure 6A) and the Complex Medically Relevant Gene regions ([39, 40]) as depicted in Figure 6B. Figure 6 illustrates that alignments obtained using personalized references consistently achieve the most even allelic balance at HET sites. The personalized approaches show less bias than the pangenome reference, with the linear reference exhibiting the most imbalance. The trend is particularly evident at longer indels (left and right extremes), where impute-first references maintain a more even allelic balance compared to the others.

2.3.3 Variant calling accuracy

Using RTG tools [35] and the GIAB HG001 truth VCF, we measured variant-level precision, recall and F1 with respect to the HG001 dataset. Measurements using Giraffe together with a linear or pangenome reference are labeled as "Giraffe (linear)" and "Giraffe (pangenome)" respectively. Measurements using impute-first workflows are labeled with "Giraffe*" together with a designator (rgc1, rgc5, bbgc5, or bbbc5).

The impute-first workflows consistently achieved greater precision and recall than BWA-MEM as well as Giraffe when using its linear and pangenome indexes. This was true both for SNVs and for indels (Figure 7b). The result was consistent whether we measured variants only in high-confidence regions (Figure 7a & b, Figure S2), in the context of GRCh38 Complex Medically Relevant Gene (CMRG) regions (Table S3), or in the more challenging GIAB-labeled regions (MHC, allDifficultregions, allOtherDifficultregions, and alllowmapandsegdupregions) (Table S4). The result is also consistent across allele frequencies, with the rarest alleles exhibiting the largest recall gap between the impute-first workflows and the others (Table S7).

The four impute-first workflows did not differ much in their precision and recall. However, the two workflows based on Bowtie2+bcftools performed slightly better than the two workflows based on Rowbowt. This difference was small compared to differences the impute-first and non-impute-first workflows. There was no appreciable difference between the Glimpse-based versus the Beagle-based workflows.

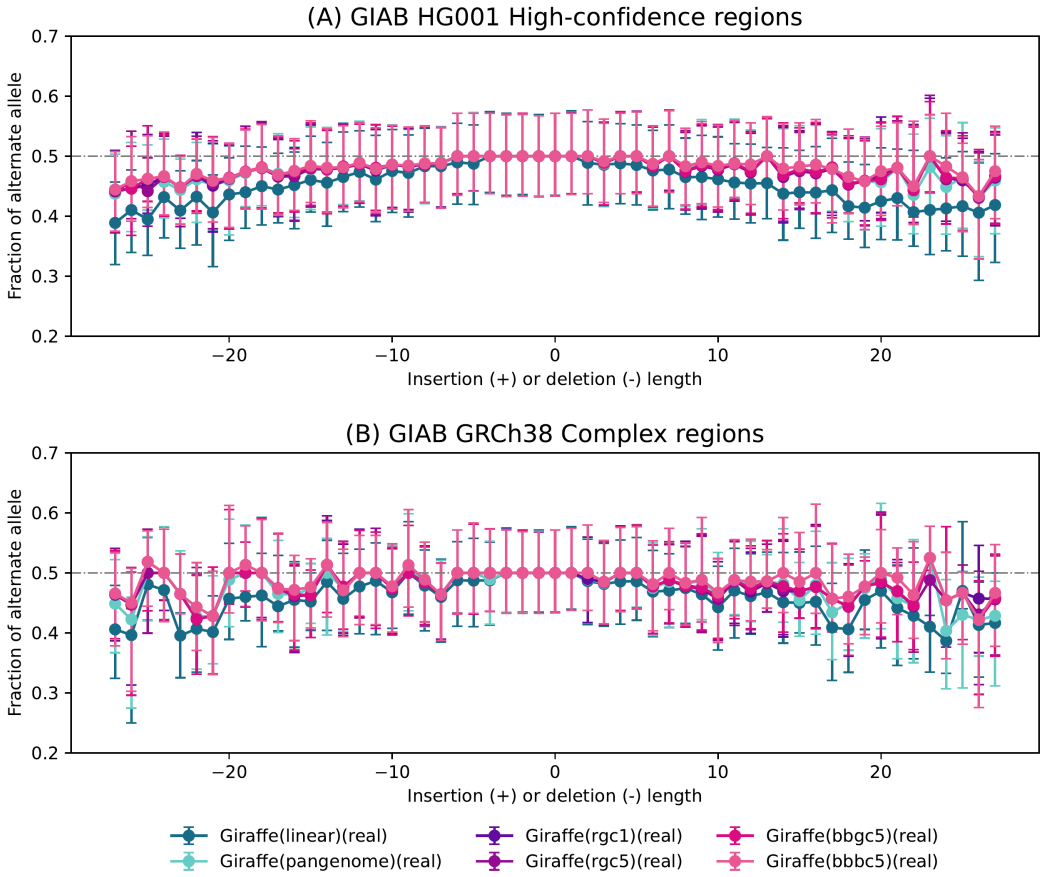
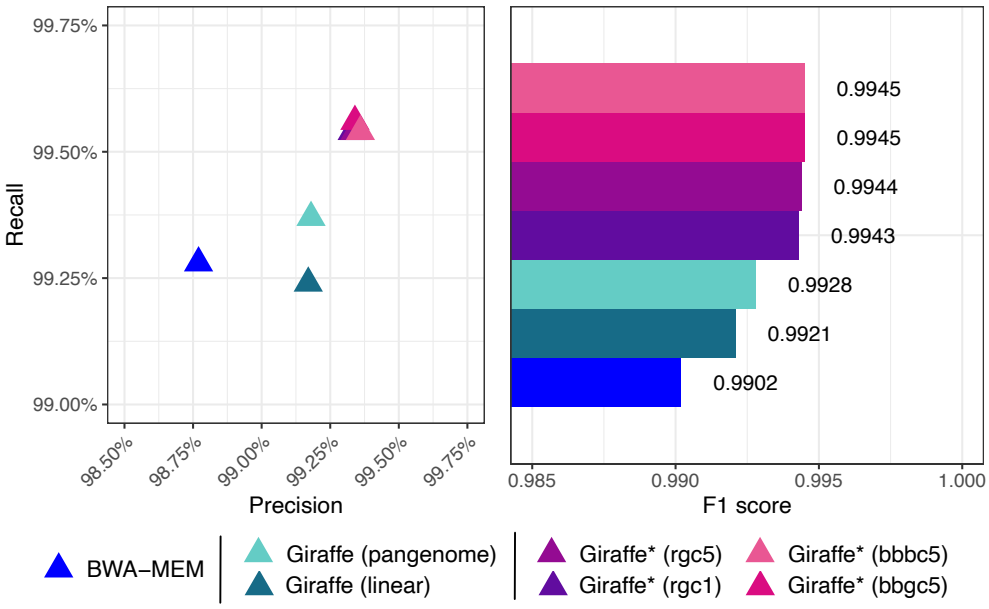


Figure 6: Bias-by-allele-length plots on HG001 donor reads generated for the 6 Giraffe-based downstream workflows. Variants are categorized by length: positive values for insertions, negative for deletions, and zero for SNVs at HET sites across the genome. (a) Allelic balance for indels and SNVs in GIAB HG001 high-confidence regions. (b) Allelic balance for indels and SNVs in GIAB GRCh38 Complex Medically Relevant Gene regions.

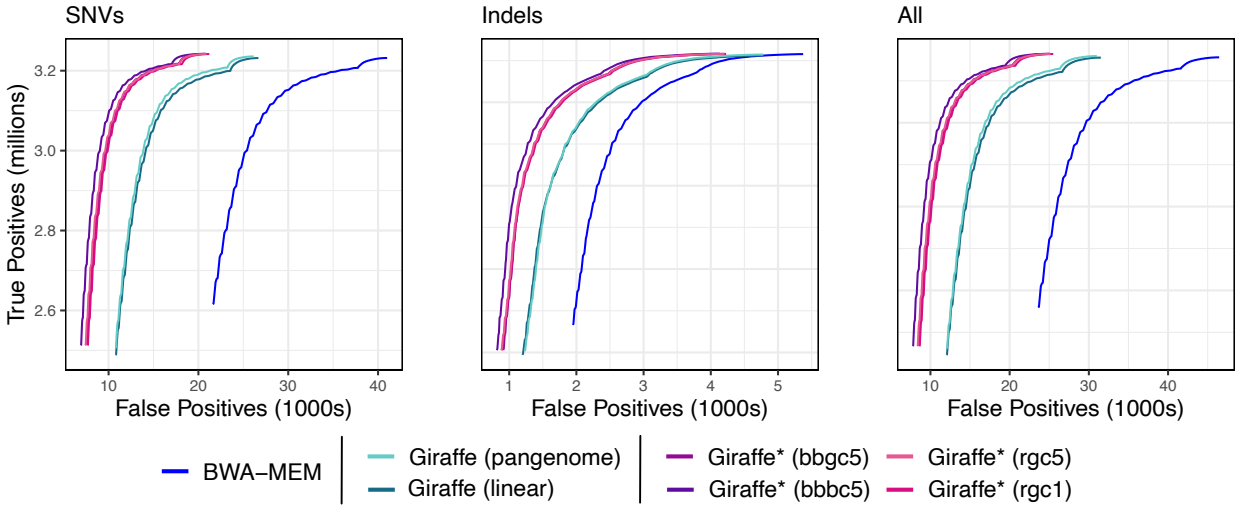
2.4 Computational efficiency

We also measured the computational overhead of all workflows, with particular attention to the overhead incurred by impute-first workflows. As noted, the impute-first workflow consists of distinct "personalization" and "downstream" phases. We evaluated the time and memory requirements for each of 6 workflows. Two of the workflows were the Giraffe linear and Giraffe pangenome workflows, neither of which involved a personalization phase. The four others were the four impute-first workflows that each involved both personalization and downstream phases. For all 6 workflows, we measured the cost of building the reference index, though it should be noted that this step can be done once in an "offline" manner for the Giraffe linear and pangenome workflows, whereas it must be done in an "online" manner for the impute-first workflows.

As shown in Figure 8b, building the Giraffe pangenome index took substantially more time than any of the other references. The impute-first references took about as much time to index as Giraffe's linear reference, though the impute-first workflows required additional time for the other persoanlization steps, i.e. genotyping and imputation. The impute-first workflow required around $20\times$ lower indexing time and $3\times$ lower memory usage compared to the Giraffe pangenome workflow (Table S5). This benefit was also reflected in the size of the graph: the impute-first workflows yielded around half the number of nodes and



(a) Precision-recall measures & F1 scores for overall variants. "Giraffe" indicates use of the standard VG Giraffe aligner. "Giraffe*" indicates use of VG Giraffe with a personalized reference produced by the personalization workflow mention in parentheses.



(b) ROC curves from `vcfeval` ROC data files for SNVs, indels and overall variants. ROC curves are stratified by variant quality value threshold. "Giraffe*" indicates use of VG Giraffe with a personalized reference produced by the personalization workflow mention in parentheses.

Figure 7: Variant calling accuracy metrics evaluated against GIAB HG001 truth VCF on High-confidence regions. HG001 donor reads were aligned with a standard linear aligner (BWA-MEM) or one of various VG Giraffe workflows. "Giraffe" indicates use of the standard VG Giraffe aligner. "Giraffe*" indicates use of VG Giraffe with a personalized reference produced by the personalization workflow mention in parentheses.

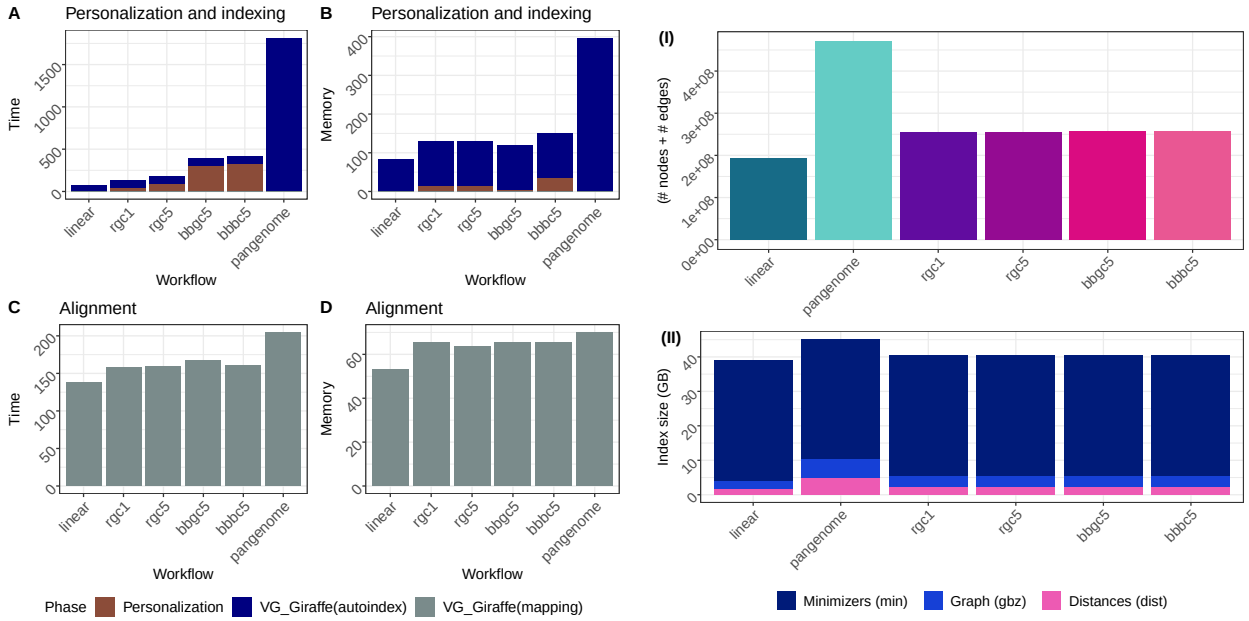
edges compared to the Giraffe pangenome (Table S6).

Figure 8 shows that, once the index has been built, the Giraffe pangenome workflow requires more time and memory than the other workflows, though these differences were smaller than the differences in indexing time and graph size. For example, the time required to align all the input reads to the Giraffe

pangenome index was 27.8% greater than the time required to align to the impute-first bbbc5 reference.

Among the impute-first combinations, the Rowbowt combinations (especially rgc1) were most efficient, combining efficient genotyping in the personalization phase with lower downstream resource requirements versus the pangenome. Overall, the impute-first based personalized references demonstrated substantially improved downstream computational performance compared to using the pangenome reference within the VG Giraffe framework. Our analysis indicates that the impute-first approach can deliver accuracy improvements while maintaining reasonable computational overhead throughout the workflow.

A key caveat is that workflows using a personalized reference necessarily require the construction of a new reference index in order to leverage the personalized calls.



(a) Time and Memory Overhead of workflows within VG Giraffe framework.

(b) Graph and index sizes of workflows generated within VG Giraffe framework.

Figure 8: Computational efficiency of VG Giraffe-based workflows, considering time, memory footprint, and index size. Compared to typical pangenome VCF usage in VG Giraffe, the personalized VCFs used in Impute-first pipelines (*rgc1, rgc5, bbgc5, bbbc5*) offered more efficient reference representation and were also comparable to the linear method in terms of computational efficiency.

3 Discussion

We introduced a practical impute-first approach for genomic analysis with the goal of minimizing reference bias, even beyond what can be achieved with pangenome references. The workflow includes initial genotyping and genotype imputation steps, which produce a diploid personalized reference. This reference can be further leveraged by indexing it and using it as a basis for downstream alignment and variant calling.

The method yields improvements in all aspects measured: higher alignment scores, improved allelic balance at heterozygous sites, and enhanced downstream variant calls. VG Giraffe alignments using an impute-first personalized reference achieved greater variant-calling accuracy VG Giraffe with a pangenome reference. Further, the VG Giraffe index built on the personalized reference was more compact, reducing the computational burdens associated with building the index and aligning reads to it.

While past methods have explored the use of personalized references for avoiding reference bias, this study is the first to demonstrate a practical workflow that (a) performs personalization and downstream analysis with the same dataset, (b) integrates imputation tools like Glimpse and Beagle 5 that work with inputs derived from (possibly low-coverage) sequencing data, (c) outputs a diploid personalized reference

genome for various downstream uses, and (d) works naturally together with a graph pangenome aligner to achieve both linear-genome-like computational overhead, and better-than-pangeome alignment accuracy.

We would note that while we concentrated on certain tools here, including BWA-MEM, VG Giraffe, Rowbowt and Glimpse, the framework we proposed is modular, and tools not assessed here could also be used. Our past work also assessed the PanGenie genotyper [41], for example, which could be integrated into the personalization phase [32].

In the future, it will be important to further improve the computational steps in the workflow's personalization phase: genotyping and imputation. For instance, the recently described Minimal Positional Substring Cover (MPSC) framework [42] can solve a version of the imputation problem in time that does not depend on the number of haplotypes in the imputation panel. It is not immediately applicable, however, as it does not currently handle phased diploid imputation.

Another avenue for improvement may be to combine results from multiple imputation tools to improve personalization accuracy. Such benefits were reported previously in ancestral DNA analysis studies [6].

While our method shows promising results with whole-genome DNA sequencing data, more work is required to establish its applicability to other assays like exome sequencing and RNA-seq. These assays exhibit uneven coverage across the genome, which complicates the initial genotyping and imputation steps. For exome sequencing, coverage is deep in exonic regions and shallow elsewhere, leading to higher genotyping and imputation accuracy in exons; accuracy will then presumably fall off in areas flanking the exons, and only low accuracy will be achieved between exons. RNA-seq poses an additional problem duew to variability in expression levels also complicates genotyping, since lowly-expressed genes may not provide sufficient data for genotyping and imputation.

We opted to use the HGSVC2 reference panel in the personalization phase of our workflow because that panel was compiled using advanced long-read assembly methods that allowed it to survey complex structural variants. However, more work is required to determine which are the most important features the imputation panel should have to minimize reference bias and maximize downstream accuracy. While we opted for a panel with relatively few individuals but a relatively wider survey of structural variants, it could also be advantageous to choose a panel with more individuals but a narrower range of SVs, such as that produced by the 1000 Genomes project. Notably, the recent Glimpse2 algorithm [43] is able to handle BioBank-scale reference panels, as demonstrated using the 150K samples in the UK BioBank.

We used the GATK HaplotypeCaller to call variants here. We did not use DeepVariant, because its internal model is already "trained" to undo the effects of reference bias, which it can readily observe in the training data that it obtains using alignment tools that produce typical levels of reference bias. More study is needed to find how the most effective tools for avoiding bias "upstream" can be combined with the most effective bias-avoiding tools "downstream." This requires tighter feedback between the two, e.g. with downstream models being trained on alignments produced by the best bias-avoiding upstream tools.

4 Methods

Figure 1 illustrates the impute-first workflow. First, a set of reads is sampled from the full set of input reads. Next, the sampled read set is used to obtain rough genotypes for the donor. The rough genotypes are passed to an imputation tool, which imputes and phases both donor haplotypes, yielding a personalized diploid reference genome. This is the "personalization" portion of the workflow, depicted in Figure 1A.

In scenarios where high-quality variant calls are desired, the workflow can be continued further downstream, as depicted in Figure 1B. In this case, the diploid personalized reference will be indexed and used to analyze the full input read set.

4.1 Personalization

4.1.1 Sampling

To minimize the computational burden of personalization phase, we tested its accuracy and efficiency when run on only a sample of the input read set. For instance, if the input reads total $30\times$ average coverage, the personalization workflow might use a random subset of those reads totaling $1\times$ average coverage. Past

work has shown that genotype imputation can be performed with excellent accuracy even when the input reads represent a low level of average coverage [36, 44].

4.1.2 Genotyping

The next step obtains a set of "rough" variant calls based on the subsampled reads. Since the sampled input might comprise a low level of average coverage, the genotypes obtained in this step are not expected to be high quality. This is acceptable in practice because the imputation step that follows combines this rough genotyping information with the large database of high-quality phased genotypes available in the imputation panel to produce an output that is much higher quality.

This step of the personalization workflow is modular; various tools could be inserted here to accomplish the rough genotyping task. In our experiments, we used four different specific toolsets, described in the following subsections.

Rowbowt Rowbowt [32] uses an r-index [12] together with an auxiliary *marker array* structure to achieve a compressed pangenome index where genotype evidence is easily tallied. First, rowbowt takes a Variant Call Format (VCF) file and a corresponding FASTA file as input, where the FASTA provides the reference sequence and the VCF describes the variants present on the haplotypes to be indexed. It then indexing the index using the prefix-free parsing algorithm [45].

We used Rowbowt to index the HGSVC2 reference panel, which comprises 34 samples, or a total of 68 haplotypes. Since some of our experiments analyzed reads from the individual with accession NA12878, we first excluded NA12878's two haplotypes before building the Rowbowt index. In addition to the HGSVC2 haplotypes, we also included the GRCh38 primary assembly as one of the indexed references. The marker array's window size w was set to 19, the value used and tested in the Rowbowt study [32]. Rowbowt computes and reports genotype likelihoods (following a method similar to past ones [46]), which can then be provided to the downstream imputation tool.

Bowtie2 + bcftools The genotyping procedure against the linear reference genome involved a series of steps. Initially, raw reads were aligned to the human reference genome, GRCh38 using Bowtie 2 v2.4.2 (Bowtie2) [29] with default parameters and 16 threads. After alignment, a pile-up of overlapping reads was generated for each genomic position using BCFtools v1.13 (`bcftools`), [47] with `bcftools mpileup` command using default settings. This pile-up information was then used for variant calling, accomplished through the `bcftools call` command with a multi-allelic calling model, retaining default parameters.

To ensure the accuracy and reliability of identified variants, the low-quality attributes were filtered out by excluding VCF records with INFO/QUAL values below 20 and INFO/DP values exceeding 100. It's important to note that both pile-up generation and variant calling stages were conducted using single-thread configurations due to the lack of multi-threading support in these respective analysis modules used.

BayesTyper BayesTyper [31] operates by matching k-mers from input reads to k-mers stored in a de Bruijn graph that represents known polymorphisms and their neighboring contexts. This evidence is then leveraged to perform genotype calls through a generative model. The genotyping process involves constructing variant graphs using alleles defined by a predetermined set of variants in a VCF file, followed by read-to-graph mapping by aligning read k-mers to graph nodes.

In our experiments, we have used BayesTyper v1.5 (BayesTyper) and built a compatible VCF file to contain all relevant variant set from the HGSVC2 reference panel. For enhanced computational efficiency, 32 threads are employed during k-mer counting and bloom filter construction, and 16 threads during the genotyping phase.

4.2 Diploid genotype imputation

4.2.1 Beagle

We used Beagle v5.1 (Beagle) [16] to phase and impute genotypes based on genotype inputs taken from the four genotyping methods. We executed Beagle with default parameters but using 32 simultaneous threads.

We performed phasing and imputation on a per-chromosome basis. For the imputation reference panel, we used the 34 samples from the HGSVC2 project, excluding NA12878 and family members (NA12889 & NA12890). The genetic map files were sourced from the Beagle web resource [48].

4.2.2 GLIMPSE

We used Glimpse v1.0.0 [17] to impute genotypes based on genotype inputs taken from the four genotyping methods. We used the same reference panel and genetic map as for the Beagle experiments, first converting the genetic map to Glimpse's format. We executed Glimpse with 32 simultaneous threads. We ran all of Glimpse's distinct modules: `chunk`, `ligate`, `phase`, and `sample` with default parameters. As we did for Beagle, we conducted the imputation on a per-chromosome basis.

4.3 Construction of personalized reference and index

4.3.1 Personalized linear reference

We used `bcftools consensus` to construct a diploid reference genomes in the FASTA format. We used `bcftools consensus` to insert the imputed, phased variants from the rgc1personalization workflow into the GRCh38 reference, creating two personalized haplotypes. We next created indexes for these haplotypes (`Haplotype_1`, `Haplotype_2`) as well as for the GRCh38 reference using BWA-MEM's (`bwa index`) command with default parameters.

4.3.2 Personalized graph reference

We generated personalized reference graphs for HG001 using `vg autoindex --workflow giraffe` from `vg` version 1.46.0. We used default parameters and 32 simultaneous threads. As inputs, we provided both the GRCh38 reference FASTA and the personalized diploid variant calls from the corresponding personalization workflow. The resulting graph was a "personalized reference graph" embedding HG001's unique variants. That is, for variants that were called homozygous ALT compared to the GRCh38 reference, those positions were switched to the ALT allele. For variants that were heterozygous, the graph will contain alternate paths for both of the two alleles. To assess size and complexity of these graphs (e.g. for Figure 8b), we converted the compressed `.gbz` indexes to `gfa` format graphs.

Additionally, for further downstream evaluations and comparisons that include index sizes and indexing performance, we performed index constructions for linear (without VCF) and pangenome (1kGP phase 3 VCF) combinations using the GRCh38 reference fasta in the `autoindex` workflow with `giraffe`.

4.4 Alignment and variant calling

4.4.1 Alignment with linear aligner

For the "alignment scores" analysis discussed in Figure 5, we used BWA-MEM with default parameters to align the full set of HG001 donor reads to the standard GRCh38 reference. We also used BWA-MEM to align all the reads to the both haplotypes of the personalized diploid reference produced using the "impute-first: rgc1" method. For the personalized reference, reads were aligned separately to both of the personalized haplotypes. After this, the alignment with the maximum scores between the two haplotypes was chosen for the analysis.

4.4.2 Alignment with VG Giraffe

Using the personalized reference graph indexes created through the VG Giraffe `autoindex` workflow, we employed the VG Giraffe mapping module (`vg giraffe`) to align the full set of HG001 donor reads. For improved accuracy, we supplied read-specific statistics to the algorithm, including the fragment length mean and standard deviation as inferred for the HG001 dataset by BWA-MEM. This alignment process resulted in standard BAM files for the respective combinations, serving as input for subsequent analyses. We followed the same alignment procedure for all the Giraffe-based workflows.

4.4.3 Variant calling with GATK HaplotypeCaller

In the variant calling phase, BAM alignment outputs from VG Giraffe workflows (linear, pangenome, impute-first combinations) were processed. After formatting the BAMs to GATK (v4.2.6.1) standards, GATK HaplotypeCaller (v2.24.1) [49] was applied to each BAM with the GRCh38 genome as the reference. The `HaplotypeCaller` identified potential variants for the HG001 from the corresponding BAMs. These results were captured in VCF files and benchmarked against the GIAB HG001 truth VCF.

4.5 Evaluation

4.5.1 Measuring genotyping and imputation accuracy

We analyzed the accuracy of the personalization workflows' diploid genotypes in two ways. First we considered allele-by-allele precision and recall, considering the alternate (ALT) allele calls from the sample-specific truth VCF to be the positive class. Specifically, every diploid genotype called by a method is considered as a pair of individual allele calls. If a given allele call is an alternate (ALT) allele and there is at least one ALT allele present in the true diploid genotype at that site, it counts as a true positive (TP). If the given allele is a reference allele (REF) and there is at least one REF allele in the true diploid genotype, this is a true negative (TN). If the given allele is an ALT but the true genotype is homozygous REF, this is counted as one false positive (FP). Finally, if the given allele is REF but the true genotype is homozygous ALT, this is counted as one false negative (FN).

Second, we considered precision and recall with respect to sites that were either truly heterozygous or called heterozygous. If a heterozygous call made by a method is truly heterozygous, this was counted as a true positive (TP). False positives, false negatives, and true negatives are defined accordingly.

In both cases, precision and recall are computed as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

4.5.2 Measuring window accuracy

To evaluate the accuracy of the short-range phasing surrounding each marker, for both haplotypes in the truth set, we extracted the 200bp flanking sequences around each polymorphic site and tallied their presence/absence in the test/impute-first genotypes. We call this "window accuracy" and calculate it as:

$$\text{Window Accuracy} = \frac{\text{Matching 200bp windows}}{\text{Total 200bp windows}}$$

4.5.3 Measuring accuracy of personalized genome alignments

For the analysis presented in Results 2.3 and Figure 5, reads aligned with BWA-MEM were selected based on variations in alignment scores across the GRCh38 and the two haplotype copies constructed from the HG001 "impute-first: rgc1" variants. Emphasis was placed on reads with non-zero score differences between the references. This subset of reads was then analyzed for both positive and negative score differences. For each read, the optimal alignment score from either of the two haplotype copies was determined and considered as the personalized score for that read. Following this, a comparison was conducted to assess the differences between the alignment scores of the GRCh38 and the personalized reference.

4.5.4 Measuring allelic balance at HETs

To quantify reference bias, we ran *biastools* including its context-aware assignment algorithm [38]. This tool examines the alignment evidence overlapping heterozygous variants. We ran *biastools* separately on per-chromosome HG001 SAM files, extracted from the workflow-generated BAM outputs. At heterozygous sites, *biastools* tallies and reports the allelic balance, i.e. the number of ALT-carrying reads overlapping the site divided by the sum of the ALT- and REF-carrying reads overlapping the site. To identify heterozygous sites, *biastools* was given the GIAB HG001 v4.2 truth VCF. We configured *biastools* to limit its analysis to

HETs contained in the GIAB-designated high-confidence regions of HG001 (Figure 6a). We repeated this, asking *biastools* to limit its analysis to HETs in the GRCh38 challenging, medically relevant genes (CMRGs) (Figure 6b). Figure 6 first partitions HET sites according to the length of the inserted or deleted sequence, with SNVs considered to be length zero. For each length category, the plot shows median allelic balance.

4.5.5 Assessing variant call accuracy

We ran `rtgtools vcfeval` using the latest HG001 v4.2 truth VCF against each of the callsets generated from GATK on feeding corresponding BAMs of linear, pangenome, impute-first workflows generated within VG Giraffe. The evaluations were performed in aggregate (overall variants) mode. For the high-confidence regions, we also stratified by SNVs versus indels.

We used the RTG Tools [35] `vcfeval` command (v3.12.1) to evaluate variant calling accuracy. Specifically, we compared variant calls made by the alternate workflows to the true genotypes reported in the GIAB HG001 v4.2 truth VCF [34]. For Figure 7 and Table S2, we limited the analysis to variants in regions called "high confidence" in HG001 by GIAB. For Table S3, we limited the analysis to the challenging, medically relevant genes (CMRGs) discussed by Wagner et al. [50]. For Table S4, we reported results for a few other challenging subsets of genomic regions, specifically the "MHC," "alldifficultregions," "allOtherDifficultregions," and "allowmapandsegdupregions" regions as defined by GIAB.

We used the default optimized threshold parameters in `rtgtools vcfeval` to summarize standard precision, recall, and F1 accuracy metrics for each callset compared to the GIAB truth. We used the ROC output from `vcfeval` to assess performance (e.g. in Figure 7b) on the high-confidence regions for SNP, indel, and overall variants.

5 Funding

NSKV, TM and BL were supported by NIH grants R01HG011392 and R35GM139602 to BL. Work was carried out at the Advanced Research Computing at Hopkins (ARCH) core facility (rockfish.jhu.edu), which is supported by the National Science Foundation (NSF) grant number OAC 1920103.

6 Availability of data and materials

- GRCh38 reference fasta
- HG001 Fastq files: SRR622457, SRR14724533
- HGSVC, phase2 haplotype reference panel
- Genetic maps
- GRCh38 phase3 1kGP callset
- GIAB HG001 v4.2.1 truth VCF & High-confidence regions
- GIAB GRCh38 v3.1 Genome stratifications
- GIAB GRCh38 Complex Medically Relevant Genes (CMRG) regions

Scripts for the experiments described in the article are at Impute-First Alignment framework.

7 Authors' contributions

NSKV, TM and BL designed the method. NSKV wrote the software and performed the experiments. NSKV and BL wrote the manuscript. All authors read and approved the final manuscript.

8 Competing interests

The authors declare that they have no competing interests.

References

1. Brandt, D. Y., Aguiar, V. R., Bitarello, B. D., Nunes, K., Goudet, J. & Meyer, D. Mapping Bias Overestimates Reference Allele Frequencies at the HLA Genes in the 1000 Genomes Project Phase I Data. *G3 (Bethesda)* **5,** 931–941 (Mar. 2015).
2. Van de Geijn, B., McVicker, G., Gilad, Y. & Pritchard, J. K. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat Methods* **12,** 1061–1063 (Nov. 2015).
3. Degner, J. F., Marioni, J. C., Pai, A. A., Pickrell, J. K., Nkadori, E., Gilad, Y. & Pritchard, J. K. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25,** 3207–3212 (Dec. 2009).
4. Rozowsky, J., Abyzov, A., Wang, J., Alves, P., Raha, D., Harmanci, A., Leng, J., Bjornson, R., Kong, Y., Kitabayashi, N., *et al.* AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol Syst Biol* **7,** 522 (Aug. 2011).
5. Salavati, M., Bush, S. J., Palma-Vera, S., McCulloch, M. E. B., Hume, D. A. & Clark, E. L. Elimination of Reference Mapping Bias Reveals Robust Immune Related Allele-Specific Expression in Crossbred Sheep. *Front Genet* **10,** 863 (2019).
6. Martiniano, R., Garrison, E., Jones, E. R., Manica, A. & Durbin, R. Removing reference bias and improving indel calling in ancient DNA data analysis by mapping to a sequence variation graph. *Genome Biol* **21,** 250 (Sept. 2020).
7. Günther, T. & Nettelblad, C. The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLoS Genet* **15,** e1008302 (July 2019).
8. Groza, C., Kwan, T., Soranzo, N., Pastinen, T. & Bourque, G. Personalized and graph genomes reveal missing signal in epigenomic data. *Genome biology* **21,** 1–22 (2020).
9. Sirén, J., Välimäki, N. & Mäkinen, V. Indexing graphs for path queries with applications in genome research. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **11,** 375–388 (2014).
10. Siren, 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.* Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science* **374,** abg8871 (Dec. 2021).
11. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature biotechnology* **37,** 907–915 (2019).
12. Kuhnle, A., Mun, T., Boucher, C., Gagie, T., Langmead, B. & Manzini, G. Efficient Construction of a Complete Index for Pan-Genomics Read Alignment. *J Comput Biol* **27,** 500–513 (Apr. 2020).
13. Garrison, E., Siren, J., Novak, A. M., Hickey, G., Eizenga, J. M., Dawson, E. T., Jones, W., Garg, S., Markello, C., Lin, M. F., *et al.* Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* **36,** 875–879 (Oct. 2018).
14. Pritt, J., Chen, N.-C. & Langmead, B. FORGe: prioritizing variants for graph genomes. *Genome biology* **19,** 220 (2018).
15. Chen, N. C., Solomon, B., Mun, T., Iyer, S. & Langmead, B. Reference flow: reducing reference bias using multiple population genomes. *Genome Biol* **22,** 8 (Jan. 2021).
16. Browning, B. L., Zhou, Y. & Browning, S. R. A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am. J. Hum. Genet.* **103,** 338–348 (Sept. 2018).
17. Rubinacci, S., Ribeiro, D. M., Hofmeister, R. J. & Delaneau, O. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nat Genet* **53,** 120–126 (Jan. 2021).

18. 1000 Genomes Project Consortium and others. A global reference for human genetic variation. *Nature* **526,** 68–74 (Oct. 2015).

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

20. Yuan, S. & Qin, Z. *Read-mapping using personalized diploid reference genome for RNA sequencing data reduced bias for detecting allele-specific expression* in *2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops* (2012), 718–724.

21. Liu, X., MacLeod, J. N. & Liu, J. iMapSplice: Alleviating reference bias through personalized RNA-seq alignment. *PloS one* **13** (2018).

22. Letcher, B., Hunt, M. & Iqbal, Z. Gramtools enables multiscale variation analysis with genome graphs. *Genome Biol* **22,** 259 (Sept. 2021).

23. Otto, T. D., Sanders, M., Berriman, M. & Newbold, C. Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology. *Bioinformatics* **26,** 1704–1707 (July 2010).

24. Turro, E., Su, S. Y., alves, Â., Coin, L. J., Richardson, S. & Lewin, A. Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome Biol* **12,** R13 (2011).

25. Yuan, S., Johnston, H. R., Zhang, G., Li, Y., Hu, Y. J. & Qin, Z. S. One Size Doesn't Fit All - RefEditor: Building Personalized Diploid Reference Genome to Improve Read Mapping and Genotype Calling in Next Generation Sequencing Studies. *PLoS Comput Biol* **11,** e1004448 (Aug. 2015).

26. Durbin, R. Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). *Bioinformatics* **30,** 1266–1272 (May 2014).

27. Li, H. *seqtk* `https://github.com/lh3/seqtk`. 2023.

28. Church, D. M., Schneider, V. A., Steinberg, K. M., Schatz, M. C., Quinlan, A. R., Chin, C. S., Kitts, P. A., Aken, B., Marth, G. T., Hoffman, M. M., *et al.* Extending reference assembly models. *Genome Biol.* **16,** 13 (Jan. 2015).

29. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9,** 357–359 (Mar. 2012).

30. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27,** 2987–2993 (Nov. 2011).

31. Sibbesen, J. A., Maretty, L. & Krogh, A. Accurate genotyping across variant classes and lengths using variant graphs. *Nat Genet* **50,** 1054–1059 (July 2018).

32. Mun, T., Vaddadi, N. S. K. & Langmead, B. Pangenomic genotyping with the marker array. *Algorithms for Molecular Biology* **18,** 1–17 (2023).

33. Poplin, R., Ruano-Rubio, V., DePristo, M., Fennell, T., Carneiro, M., Van der Auwera, G., Kling, D., Gauthier, L., Levy-Moonshine, A., Roazen, D., *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* (2018).

34. Zook, J. M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., Weng, Z., Liu, Y., Mason, C. E., Alexander, N., *et al.* Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific data* **3,** 1–26 (2016).

35. 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.* Comparing variant call files for performance benchmarking of next-generation sequencing variant calling pipelines. *BioRxiv,* 023754 (2015).

36. Li, J. H., Mazur, C. A., Berisa, T. & Pickrell, J. K. Low-pass sequencing increases the power of GWAS and decreases measurement error of polygenic risk scores compared to genotyping arrays. *Genome Res* **31,** 529–537 (Apr. 2021).

37. Chaisson, M. J., Sanders, A. D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E. J., Rodriguez, O. L., Guo, L., Collins, R. L., *et al.* Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature communications* **10,** 1784 (2019).

38. Lin, M.-J., Iyer, S., Chen, N.-C. & Langmead, B. Measuring, visualizing and diagnosing reference bias with biastools. *bioRxiv,* 2023–09 (2023).

39. Wagner, J., Olson, N. D., Harris, L., McDaniel, J., Cheng, H., Fungtammasan, A., Hwang, Y. C., Gupta, R., Wenger, A. M., Rowell, W. J., *et al.* Curated variation benchmarks for challenging medically relevant autosomal genes. *Nat Biotechnol* **40,** 672–680 (May 2022).

40. NIST. *Medically Relevant Genes* [Online]. Available from: `https://github.com/usnistgov/cmrg-benchmarkset-manuscript/tree/master/data/gene_coords/unsorted/GRCh38_mrg_full_gene.bed`. 2023.

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

42. Sanaullah, A., Zhi, D. & Zhang, S. Minimal positional substring cover is a haplotype threading alternative to Li and Stephens model. *Genome Res* **33,** 1007–1014 (July 2023).

43. Rubinacci, S., Hofmeister, R. J., Sousa da Mota, B. & Delaneau, O. Imputation of low-coverage sequencing data from 150,119 UK Biobank genomes. *Nat Genet* **55,** 1088–1090 (July 2023).

44. Martin, A. R., Atkinson, E. G., Chapman, S. B., Stevenson, A., Stroud, R. E., Abebe, T., Akena, D., Alemayehu, M., Ashaba, F. K., Atwoli, L., *et al.* Low-coverage sequencing cost-effectively detects known and novel variation in underrepresented populations. *Am J Hum Genet* **108,** 656–668 (Apr. 2021).

45. Boucher, C., Gagie, T., Kuhnle, A., Langmead, B., Manzini, G. & Mun, T. Prefix-free parsing for building big BWTs. *Algorithms Mol Biol* **14,** 13 (2019).

46. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27,** 2987–2993 (Nov. 2011).

47. 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.* Twelve years of SAMtools and BCFtools. *Gigascience* **10** (Feb. 2021).

48. Beagle. *Genetic Maps* [Online]. Available from: `https://bochet.gcc.biostat.washington.edu/beagle/genetic_maps/`. 2023.

49. Van der Auwera, G. A. & O'Connor, B. D. *Genomics in the cloud: using Docker, GATK, and WDL in Terra* (O'Reilly Media, 2020).

50. Wagner, J., Olson, N. D., Harris, L., McDaniel, J., Cheng, H., Fungtammasan, A., Hwang, Y. C., Gupta, R., Wenger, A. M., Rowell, W. J., *et al.* Curated variation benchmarks for challenging medically relevant autosomal genes. *Nat Biotechnol* **40,** 672–680 (May 2022).

Supplementary Figures

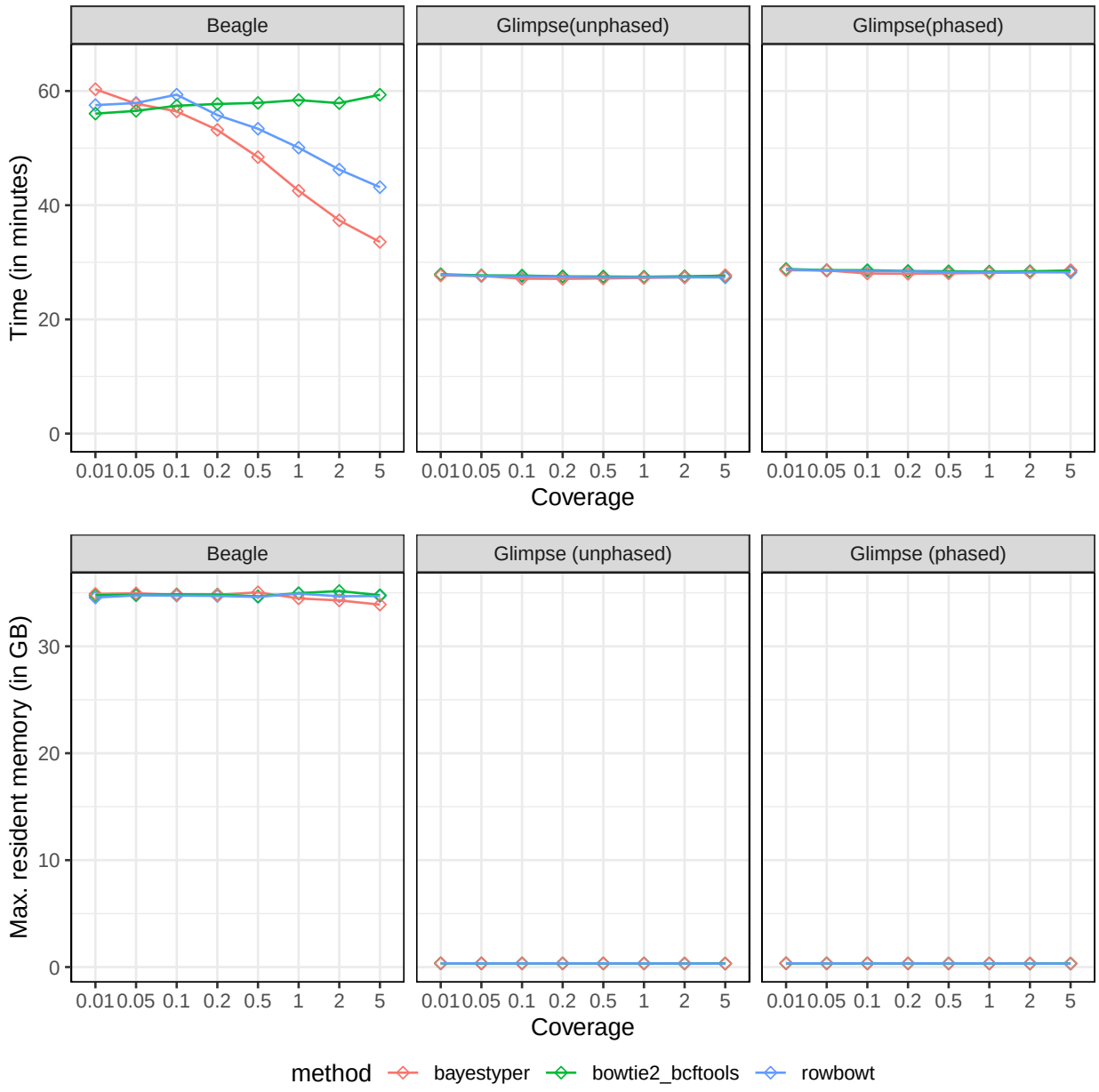


Figure S1: Computational overhead for Beagle and Glimpse stratified by the genotyping tool used upstraem, and as a function of the read coverage provided to the genotyper. Measurements for Glimpse were taken in both its phased and unphased modes.

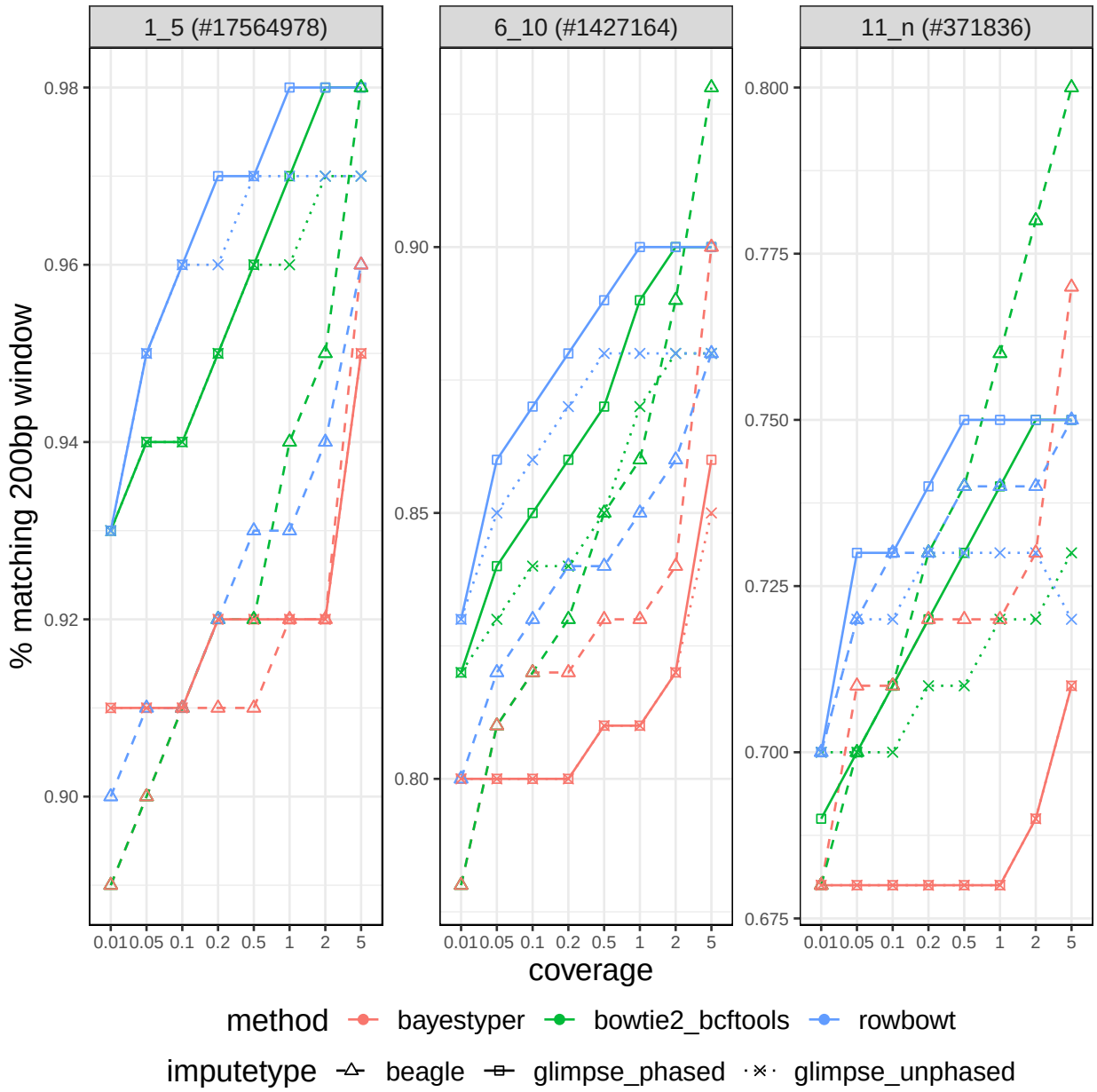


Figure S2: Window Accuracy for Beagle and Glimpse (in both phased and unphased modes) for different methods on diverse coverages.

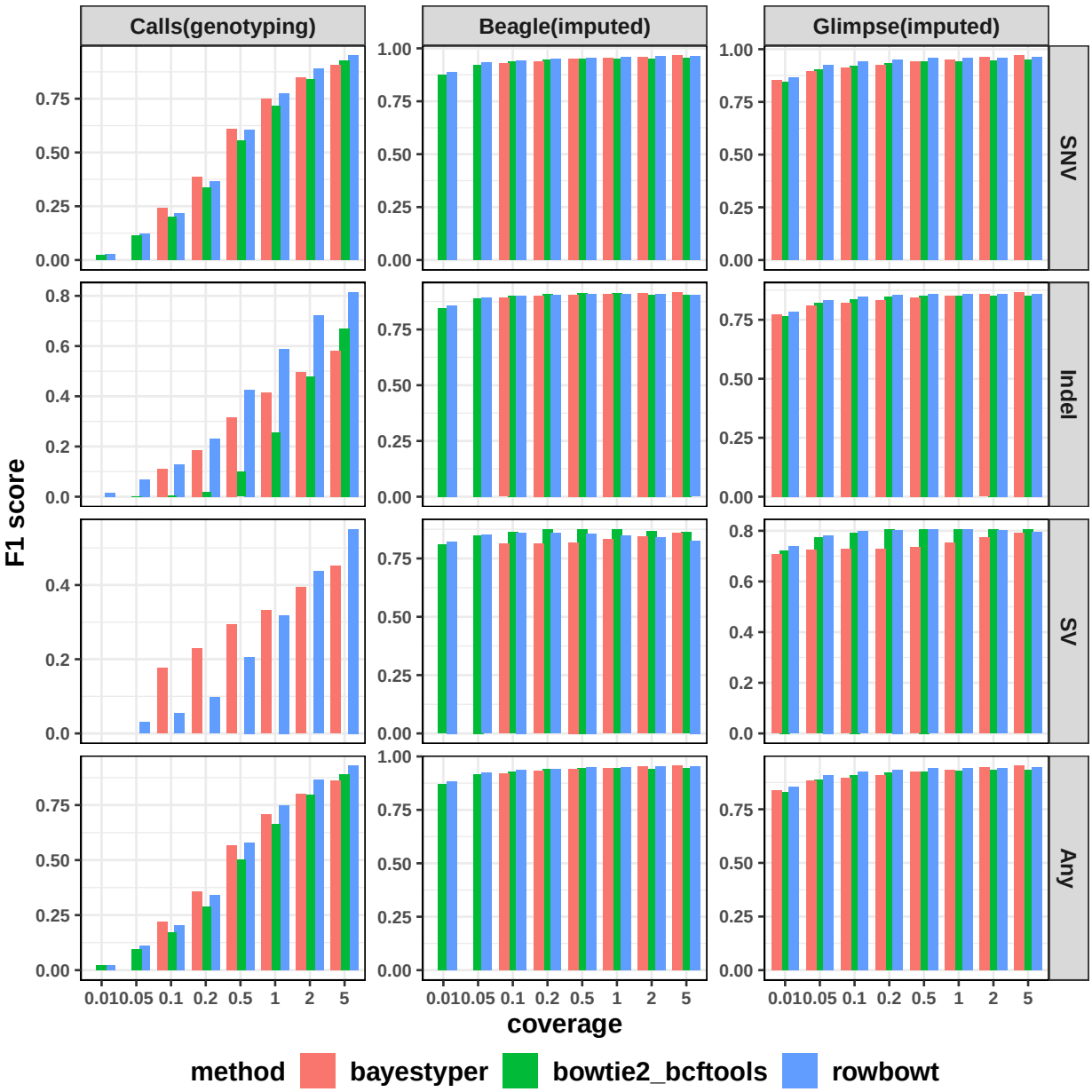


Figure S3: F1 score of alternate allele calls (ALT), stratified by variant type, for the personalized genomes constructed using each alignment/genotyping method in the Impute-first alignment workflow.

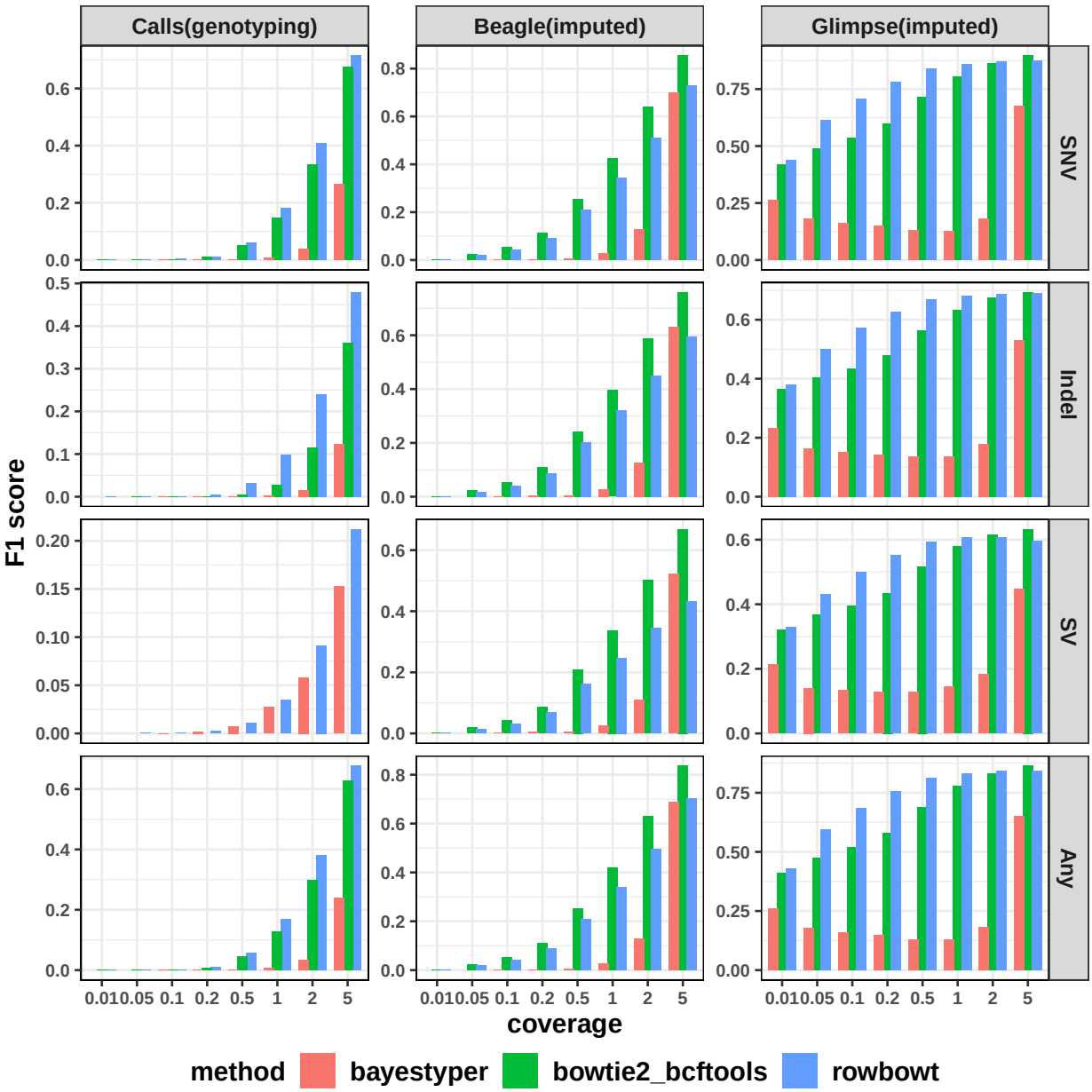


Figure S4: F1 score of heterozygous calls (HET), stratified by variant type, for the personalized genomes constructed using each alignment/genotyping method in the Impute-first alignment workflow.

Supplementary Tables

Table S1: Alignment score comparison between the `rgc1`-imputed diploid personalized reference and the standard linear GRCh38 reference.

Alignment Score Difference (Personalized - GRCh38)	**Read Count**	**Fraction (%)**	**Mean**	**Median**	**Variance**	**Std. Deviation**
Positive	50,065,960	96.71%	7.70	5.00	100.97	10.05
Negative	1,702,409	3.29%	-7.58	-5.00	112.81	10.62

Table S2: Variant calling performance metrics for HG001 real donor reads, stratified by SNVs, indels, and overall variants, across different reference combinations within GIAB HG001 high-confidence regions.

Mode	Workflow	Precision (%)	Recall (%)	F1 score
All	BWA-MEM	98.77	99.28	0.9902
	VG Giraffe (linear)	99.17	99.24	0.9921
	VG Giraffe (pangenome)	99.18	99.37	0.9928
	VG Giraffe* (rgc1)	99.33	99.54	0.9943
	VG Giraffe* (rgc5)	99.34	99.54	0.9944
	VG Giraffe* (bbgc5)	99.34	99.56	0.9945
	VG Giraffe* (bbbc5)	99.36	99.54	0.9945
SNV	BWA-MEM	98.77	99.29	0.9903
	VG Giraffe (linear)	99.21	99.28	0.9925
	VG Giraffe (pangenome)	99.22	99.41	0.9932
	VG Giraffe* (rgc1)	99.37	99.60	0.9949
	VG Giraffe* (rgc5)	99.38	99.60	0.9949
	VG Giraffe* (bbgc5)	99.38	99.63	0.9951
	VG Giraffe* (bbbc5)	99.41	99.60	0.9950
Indel	BWA-MEM	98.92	98.95	0.9894
	VG Giraffe (linear)	99.06	98.80	0.9893
	VG Giraffe (pangenome)	99.06	98.90	0.9898
	VG Giraffe* (rgc1)	99.18	98.96	0.9907
	VG Giraffe* (rgc5)	99.16	98.99	0.9908
	VG Giraffe* (bbgc5)	99.18	98.97	0.9908
	VG Giraffe* (bbbc5)	99.18	99.02	0.9910

Table S3: Variant calling performance metrics for HG001 real donor reads for overall variants, across different reference combinations within GIAB GRCh38 Complex Medically Relevant Gene (CMRG) regions.

Workflow	**Precision (%)**	**Recall (%)**	**F1 score**
BWA-MEM	85.37	97.87	0.9120
VG Giraffe (linear)	85.57	97.81	0.9128
VG Giraffe (pangenome)	85.58	98.03	0.9138
VG Giraffe* (rgc1)	85.55	98.22	0.9144
VG Giraffe* (rgc5)	85.57	98.21	0.9146
VG Giraffe* (bbgc5)	85.60	98.24	0.9148
VG Giraffe* (bbbc5)	85.77	98.19	0.9156

Table S4: Variant calling performance metrics for HG001 real donor reads for overall variants, across different reference combinations within GIAB GRCh38 stratifications.

Region	Workflow	Precision (%)	Recall (%)	F1 score
MHC	BWA-MEM	81.13	92.52	0.8646
	VG Giraffe (linear)	83.13	92.28	0.8746
	VG Giraffe (pangenome)	82.03	93.02	0.8718
	VG Giraffe* (rgc1)	84.63	92.59	0.8843
	VG Giraffe* (rgc5)	84.90	92.58	0.8858
	VG Giraffe* (bbgc5)	84.72	92.99	0.8867
	VG Giraffe* (bbbc5)	88.77	91.80	0.9026
alldifficultregions	BWA-MEM	56.08	91.45	0.6952
	VG Giraffe (linear)	59.21	91.72	0.7196
	VG Giraffe (pangenome)	59.39	92.13	0.7222
	VG Giraffe* (rgc1)	59.99	93.15	0.7298
	VG Giraffe* (rgc5)	60.04	93.13	0.7301
	VG Giraffe* (bbgc5)	60.00	93.28	0.7303
	VG Giraffe* (bbbc5)	60.25	93.17	0.7318
allOtherDifficultregions	BWA-MEM	18.08	82.75	0.2967
	VG Giraffe (linear)	23.28	82.59	0.3633
	VG Giraffe (pangenome)	23.20	84.37	0.3639
	VG Giraffe* (rgc1)	24.65	85.32	0.3824
	VG Giraffe* (rgc5)	24.71	85.28	0.3832
	VG Giraffe* (bbgc5)	24.70	85.62	0.3834
	VG Giraffe* (bbbc5)	24.98	84.25	0.3854
alllowmapandsegdupregions	BWA-MEM	40.27	81.74	0.5395
	VG Giraffe (linear)	47.10	83.40	0.6020
	VG Giraffe (pangenome)	48.27	84.21	0.6137
	VG Giraffe* (rgc1)	49.98	88.00	0.6375
	VG Giraffe* (rgc5)	50.06	87.90	0.6379
	VG Giraffe* (bbgc5)	50.03	88.46	0.6391
	VG Giraffe* (bbbc5)	50.33	88.26	0.6410

Table S5: Time and memory taken by each VG Giraffe-based workflow in the personalization, downstream indexing, and downstream alignment steps.

Phase	Workflow	Time (min)	Memory (GB)
Personalization	Impute-first: rgc1	50.23	14.45
	Impute-first: rgc5	94.97	14.45
	Impute-first: bbgc5	298.36	3.44
	Impute-first: bbbc5	329.17	34.78
Giraffe autoindex	linear	76.735	84.366
	pangenome	1817.522	397.303
	Impute-first: rgc1	86.483	113.902
	Impute-first: rgc5	91.604	114.077
	Impute-first: bbgc5	91.910	114.783
	Impute-first: bbbc5	87.335	115.220
Giraffe alignment	linear	137.529	53.062
	pangenome	204.824	70.086
	Impute-first: rgc1	157.973	65.249
	Impute-first: rgc5	158.795	63.799
	Impute-first: bbgc5	167.117	65.432
	Impute-first: bbbc5	160.250	65.440

Table S6: Index and graph size measurements for the VG Giraffe graphs generated in various workflows.

Input to `autoindex`	`.gbz` file (GB)	`.dist` file (GB)	`min` file (GB)	Total (GB)	Size (nodes + edges)
Linear GRCh38	2.6	1.5	35	39.1	193,945,149
1000 Genomes phase-3 pangenome	5.4	4.8	35	45.2	469,910,808
rgc1-imputed diploid	3.1	2.3	35	40.4	255,676,786
rgc5-imputed diploid	3.1	2.3	35	40.4	255,645,758
bbgc5-imputed diploid	3.1	2.3	35	40.4	256,688,029
bbbc5-imputed diploid	3.1	2.3	35	40.4	258,286,403

Variant Frequency	Workflow	Precision (%)	Recall (%)	F1 score
MAF $\geq$ 5% (n = 1,164,564)	BWA-MEM	99.77	99.67	0.9972
	VG Giraffe (linear)	99.81	99.69	0.9975
	VG Giraffe (pangenome)	99.84	99.78	0.9981
	VG Giraffe* (rgc1)	99.86	99.78	0.9982
	VG Giraffe* (rgc5)	99.86	99.78	0.9982
	VG Giraffe* (bbgc5)	99.86	99.79	0.9982
	VG Giraffe* (bbbc5)	99.86	99.78	0.9982
0.5% $\leq$ MAF $<$ 5% (n = 171,987)	BWA-MEM	99.96	99.87	0.9992
	VG Giraffe (linear)	99.96	99.89	0.9993
	VG Giraffe (pangenome)	99.97	99.92	0.9994
	VG Giraffe* (rgc1)	99.97	99.92	0.9995
	VG Giraffe* (rgc5)	99.97	99.93	0.9995
	VG Giraffe* (bbgc5)	99.97	99.93	0.9995
	VG Giraffe* (bbbc5)	99.97	99.92	0.9995
MAF $<$ 0.5% (n = 58,503)	BWA-MEM	99.98	99.71	0.9984
	VG Giraffe (linear)	99.99	99.76	0.9988
	VG Giraffe (pangenome)	99.99	99.78	0.9988
	VG Giraffe* (rgc1)	99.99	99.84	0.9992
	VG Giraffe* (rgc5)	99.99	99.85	0.9992
	VG Giraffe* (bbgc5)	99.99	99.83	0.9991
	VG Giraffe* (bbbc5)	99.99	99.83	0.9991

Table S7: Variant calling performance metrics for all variants present in HG001 across different minor allele frequency (MAF) ranges.