Population-specific genome graphs improve high-throughput sequencing data analysis: A case study on the Pan-African genome

H. Serhat Tetikol[1,*,†], Kubra Narci[1,†], Deniz Turgut[1,†], Gungor Budak[1,†], Ozem Kalay[1], Elif Arslan[1], Sinem Demirkaya-Budak[1], Alexey Dolgoborodov[1], Amit Jain[1], Duygu Kabakci-Zorlu[1], Richard Brown[1], Vladimir Semenyuk[1], and Brandi Davis-Dusenbery[1]

[1]Seven Bridges Genomics, Charlestown, MA, USA
[†]These authors contributed equally to this work.
[*]serhat.tetikol@sevenbridges.com (corresponding author)

ABSTRACT

Graph-based genome reference representations have seen significant development, motivated by the inadequacy of the current human genome reference for capturing the diverse genetic information from different human populations and its inability to maintain the same level of accuracy for non-European ancestries. While there have been many efforts to develop computationally efficient graph-based bioinformatics toolkits, how to curate genomic variants and subsequently construct genome graphs remains an understudied problem that inevitably determines the effectiveness of the end-to-end bioinformatics pipeline. In this study, we discuss major obstacles encountered during graph construction and propose methods for sample selection based on population diversity, graph augmentation with structural variants and resolution of graph reference ambiguity caused by information overload. Moreover, we present the case for iteratively augmenting tailored genome graphs for targeted populations and test the proposed approach on the whole-genome samples of African ancestry. Our results show that, as more representative alternatives to linear or generic graph references, population-specific graphs can achieve significantly lower read mapping errors, increased variant calling sensitivity and provide the improvements of joint variant calling without the need of computationally intensive post-processing steps.

Introduction

The development of the human genome reference[1,2], which constitutes a linear sequence of nucleotides, facilitated a significant leap forward in our understanding of the human DNA, enabling the development of high-throughput sequencing technologies and consequently a plethora of genomic studies in both academic and clinical contexts[3,4]. Many of the state-of-the-art bioinformatics methods indeed rely on this high fidelity haplotype in order to make sense of the raw data outputted by sequencers. This is commonly achieved by aligning short sequencing reads against the linear reference (*alignment*) and then identifying the differences in the data with respect to the reference (*variant calling*)[5]. Subsequently, this secondary analysis stage is followed by tertiary analyses such as variant effect prediction on protein transcription[6], genome wide association studies[4,7–9] and population genetic studies[8,10,11], the validity and effectiveness of which are fundamentally determined by the genome reference used.

Even though significant efforts have been made to ensure the quality and fitness of the current human genome reference for genomic studies[12], it is derived from a handful of individuals with about 70% pertaining to a single individual, and therefore it cannot capture the genetic diversity of the vast majority of the human populations around the globe[13–15]. This issue has been highlighted by many studies over the past decade, discussing the limitations of the latest human reference genome versions GRCh37 and GRCh38 in addition to those of a linear reference in general[15–20]. Various methods for incorporating a wider breadth of genetic information into the reference genome have been proposed including nucleotide additions and extensions to the current reference[21–23], de novo assembly of raw read data to generate a population-specific consensus sequence[24–26], and graph-based references capable of simultaneously representing multiple diverse populations[27–32]. All of these methods have trade-offs between accuracy, efficiency and applicability[33]. In addition to developing a suitable data structure and appropriate algorithms to work with it, choosing the appropriate variation information to incorporate into the reference is an important but understudied problem without a straightforward solution[34].

Another level of difficulty is introduced when the execution of bioinformatics solutions in real life scenarios is considered. Most large scale sequencing studies span a long period of time with different stages of the project such as patient enrollment

and DNA sequencing being executed in multiple phases distributed over several years[35–38]. This introduces a cyclic nature to the project where, in each cycle, the sequenced samples are processed and potentially put to use in tertiary analyses without waiting for the next cycle (see Figure 1). There are both advantages and disadvantages to processing the data collected in previous cycles without waiting for the next cycle. On one hand, it is possible to produce valuable insights from the already available data, or identify potential problems in the sequencing and secondary analysis pipeline before the next batch of samples are processed. On the other hand, commonly used bioinformatics methods such as joint variant calling[39] may require the entire set of samples to produce accurate results. Another constraint in large-scale projects is that it is not feasible to switch to a newer and more accurate version of the reference genome throughout the project due to problems that can arise regarding the incompatibility of the genome coordinate system and variant harmonization.

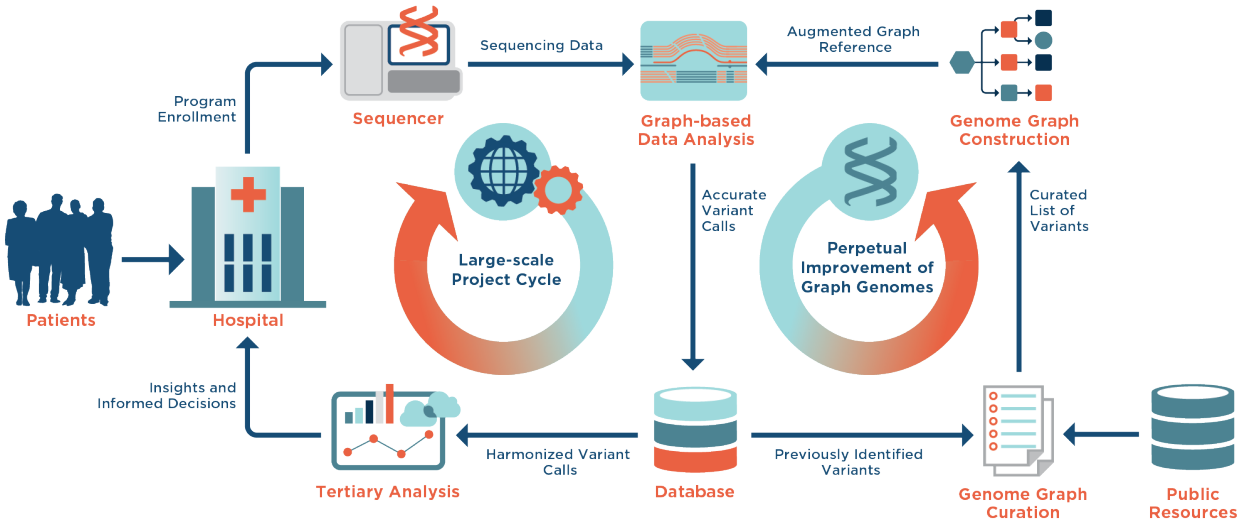

Figure 1. Steps involved in a multi-phase sequencing project. Large-scale sequencing projects are commonly executed in multiple phases, each of which comprises the sequencing and bioinformatics analysis of only a subset of the samples that are planned to be sequenced throughout the project (*Large-scale Project Cycle*). This iterative nature provides the opportunity to produce genomic information in each cycle that can be used to improve the bioinformatics processes during the project (*Perpetual Improvement of Graph Genomes*). Graph-based secondary analysis approaches can utilize this information to improve the variant detection power for the following cycles.

An optimal genome reference addressing all of these issues should at least suffice the following criteria:

- **Accurate representation of diverse genetic information**: One of the fundamental limitations of the current human genome reference is that it comprises the genomes of only a few individuals and therefore under-represents most of human populations. An ideal reference should be able to capture the genetic diversity of an entire population and/or species[13–15].

- **Compatibility with upstream and downstream methods**: To ensure the sustainability and long-term utility of the results obtained from sequencing data, the bioinformatics community has developed standards and best practices around inputs/output formats, data storage formats and workflow descriptions[40–43]. It is crucial for new methods and references to comply with these standards to maintain the longevity and compatibility of existing results.

- **Aptness for improvements and other modifications**: It is expected that future studies will identify shortcomings of the current reference and/or discover valuable genetic information that should be incorporated into the reference. The chosen genome reference should gracefully lend itself to improvements and other updates so that the downstream tools do not break and the new results can be processed together with older results without any loss of functionality. It may even be desirable that such changes to the reference are made while the project is in progress to incrementally improve it with new findings or to compensate for a deficiency. Since many sequencing projects have multiyear plans with prospective follow-up studies extending further into the future, it is very likely that the opportunity to improve the reference will present itself[35–38].

- **Computational efficiency and scalability**: Even though improvements in hardware have increased the available computational capabilities, large infrastructure investments and the incentive to use existing but older hardware remain a prohibitive barrier in front of experimenting with new bioinformatics technologies. Therefore, the tools and algorithms that can make use of the genome reference should be available (or at least theoretically possible) and computationally efficient compared to other state-of-the-art approaches such as BWA-MEM[44] for alignment and GATK[39] for variant calling.

- **Tailorability for targeted applications**: It would be unreasonable to expect that a single genome reference will be suitable for all types of genomic applications. In order to develop a reference that is targeted to an application or a population, the reference must be adaptable to a specific genomic context. The methods for constructing such references should also be available, computationally feasible and easy-to-use.

In this study, we propose a population-specific graph construction method that can meet all of the above criteria and compare its utility in next-generation sequencing (NGS) read alignment and variant calling to other approaches based on either the linear genome reference or a generic graph reference that is not tailored to the population. We show that a population-specific genome graph can significantly improve both read alignment and variant calling accuracy while being computationally efficient. Moreover, we show that genome graphs can be augmented, simultaneous with the project execution, by incorporating recently discovered genomic information to further improve the detection power of both short variants (SNPs and INDELs) and structural variants (SVs). We compare our results with those obtained by using joint variant calling[39], which is the state-of-the-art method for processing a large number of samples, and show that a graph-based approach provides most of the improvements provided by joint calling. To demonstrate the aforementioned capabilities, we choose the African population for this case study, as it is the most genetically diverse and also the most under-represented population with respect to the current human genome reference[7,25].

The accuracy of a graph-based bioinformatics pipeline is fundamentally tied to the graph reference being used with the pipeline. Therefore, in this study, we present a novel graph curation and construction method that can act as a guideline for creating a representative, accurate and efficient reference for any population of choice. The relationship between the representativeness of a graph and the diversity of the population is established. We further elaborate on the number of samples used to construct the graph reference, the sampling of the population and how they relate to the captured genetic diversity in the graph reference. We also explore key pitfalls of graph construction that can severely deteriorate the performance of the graph-based pipeline and show that our proposed method can circumvent those issues. Equally importantly, the proposed methods are not pertinent only to the specific graph-based tools used in this study but, in principle, can be applied to any other graph-based bioinformatics toolkits for genetically representative and computationally performant graph reference construction.

Results

In order to test the hypothesis that the bioinformatics analysis can be improved by utilizing a graph-based method and constructing a population-specific graph, we use the Seven Bridges GRAF pipeline[30] and benchmark the pipeline on the Illumina sequencing data of the African samples from the 1000 Genomes Project[7,45]. We split the samples into two sets, each containing the same ratio of males/females and the 7 African populations in the 1000 Genomes dataset. The *construction set* is used for graph reference construction and the *benchmarking set* is used to measure the performance of various flavors of graph references. We categorize graph references into three main types depending on the sources used to construct them.

1. **Pan-genome graph:** A graph reference that contains information from all populations around the world. This type of graph is not specific to any particular population. It can provide an overall accuracy improvement over linear references when analyzing any sample regardless of their ancestry. Pan-genome graphs are relatively easier to construct because of the abundant data in comprehensive and easily accessible public databases such as gnomAD[46] and UK BioBank[36].

2. **Population-specific graph:** A graph reference that contains genetic information pertaining to a single population. Population-specific graphs are tailored towards individuals with a common ancestry and therefore sharing a significant portion of their genetic variations. Such a graph allows the incorporation of more variants related to the population while leaving out many variants that are frequently observed in the larger human population but irrelevant to the targeted population. Many public databases are available to construct population-specific graphs. However, even though they provide good coverage over continental super-populations, specific information about a given subpopulation is usually missing.

3. **Cohort-specific graph:** A graph reference constructed using a subset of the cohort on which the graph is intended to be used. By pre-processing a subset of the sample set under study, it is possible to get an accurate description of the genetic

composition which, in turn, can be used to construct a more representative graph as opposed to relying on public data sources. More often than not, public resources may not properly cover the target population, in which case this type of graph reference is the only possible solution.

In this study, we construct all three types of graphs for the African population and compare the alignment and variant calling performance on the benchmarking set. We show that, as the graph reference becomes more tailored to the population (from type 1 to type 3 above), alignment error rates are reduced significantly and the rate of informative reads increases. In turn, the sensitivity of the pipeline improves leading to better detection of both high and low frequency variants throughout the whole genome as well as the functional regions. Including the relevant population diversity in the reference and thus making it available for read alignment and variant calling enables more consistent genotyping across all samples without the need for a joint-calling post-processing step.

Iterative analysis of population sequencing data

To simulate a multi-phase sequencing project, the construction set is split into 5 equal sized cohorts of 104 samples, leaving 141 samples for the benchmarking set. Initially, a population-specific graph is constructed using the public database gnomAD[46], which is used to process the first cohort, as shown in Figure 2. This recently expanded database covers all of the subpopulations sequenced in the 1000 Genomes project and therefore constitutes a good starting point for graph construction. The variant calls on the first cohort are combined with the public resources to generate the next graph reference which, in turn, is used to process the second cohort. This iterative procedure illustrates the nature of a multi-phase sequencing project where the graph reference is augmented with new variant calls after each iteration.

The graphs produced at each step are used to process the benchmarking set and evaluate the performance. Additionally, a pan-genome graph, which is constructed from multiple public resources and contains genetic information of many populations[30], is used to compare a non-specific graph to the population-specific graphs. Finally, all graph approaches are compared to the standard linear approach based on BWA-MEM and GATK to establish a baseline for a reliable comparison with the existing technologies. The linear BWA+GATK pipeline utilizes joint calling and VQSR variant filtration which are the recommended methods for processing a large number of whole genome samples[45].

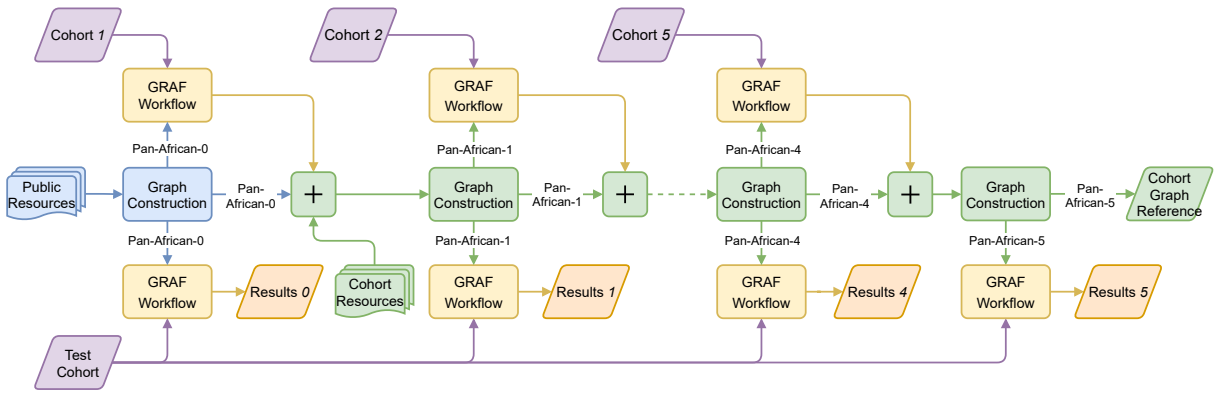

Figure 2. Iterative population-specific graph construction workflow. The initial population-specific graph reference (Pan-African 0) is constructed using publicly available variant databases. At each iteration, a subset of the population (construction set) is processed with the current graph and the variant calls are used to construct the next graph. This process is repeated until the entire construction set is exhausted. All graph references are tested on the benchmarking set and their performance is evaluated. The population-specific graphs (Pan-African 0-5) are also compared to a generic graph (Pan-Genome) containing the genetic information of many populations and a linear approach using only GRCh38 reference.

Variant Curation for Population-Specific Graphs

Variant selection for graph references still remains an open question. Previous studies mostly relied on simple heuristics or methods that may not scale well with large and/or missing information[34]. Here, we propose a framework that relies on two basic measurements on a given population: Nucleotide *diversity* within the population and absolute *divergence* from the current human genome reference, in this case, GRCh38 (see Methods)[47]. In short, nucleotide diversity measures the average genetic distance between the individuals from the same population while absolute divergence measures how distant the population is from GRCh38.

It is crucial to make the distinction between these two genetic distance measures. Graph references fundamentally rely on pre-existing genetic information which is used to construct them in the first place. The more divergent the population is from GRCh38, the more likely it is that the current bioinformatics methods have failed and will fail to discover important genetic variation. This essentially determines the detection power of a graph-based method and how much novel discovery can be expected from it. In instances of large divergence, it may be desirable to additionally incorporate results from orthogonal technologies (e.g. long-read sequencing data[48–50], which is demonstrated in this study) to compensate for the shortcomings of the technologies already used. On the other hand, a high diversity implies that the individuals from the same population are genetically distant from each other. This determines the number of samples required to construct a representative graph; diverse populations will require a larger number of samples for graph construction. This is further elucidated in this section.

To quantify these two metrics and provide a guideline for population-specific graph construction, we measure the diversity and the divergence in both the whole genome and the whole exome regions for all five populations in the 1000 Genomes dataset, as shown in Figure 3. The African population shows the highest diversity and also the largest divergence from GRCh38. This implies that the accuracy of bioinformatics tools will suffer the most when processing the sequencing data for individuals of African descent. The American population is also highly admixed compared to European, South Asian and East Asian populations, indicated by its large diversity. This is mainly caused by the large variation in the nucleotide diversities between American subpopulations (see Supplementary Material S8 for measurements on individual sub-populations). Interestingly, the East Asian population shows a large divergence from GRCh38 while showing the least amount of diversity among the five populations. This means that, even though standard linear methods are likely to miss important variation information for East Asian populations, a relatively small number of samples should be sufficient to construct a graph reference reasonably representative to the extent that can be expected by relying on the data obtained with linear methods. It is shown in the following sections that long-read sequencing data can be incorporated into graph references to account for large divergence from GRCh38. Finally, it is observed that the European population has the lowest divergence, which is expected considering the sequencing data used to construct the current human genome reference. However, the diversity within the European populations is still large enough, in fact larger than East Asian populations, to warrant the development of a population-specific graph reference. As expected, in the whole exome, both the nucleotide diversity and the absolute divergence are much smaller compared to the whole genome, nevertheless, the trends between populations remain the same.

Next, we investigate the influence of the number of samples on the representativeness of the constructed graph reference and show that it correlates well with the population's nucleotide diversity. In this experiment, we pick samples from a given population one by one and measure how much of the population's genetic variation is correctly captured in the graph reference. To include only common variants in the graph, we use an allele frequency (AF) cut-off of 5%, variants below which are discarded. At each step, we compare the content of the graph with the complete population information (obtained from all available samples), label any high ($AF \geq 5\%$) and low ($AF < 5\%$) frequency variants in the graph as true and false positives, respectively, and finally calculate true positive rate (TPR) and false positive rate (FPR). This procedure is repeated for each population independently, as shown in Figure 4 (for results with other AF threshold values, see Supplementary Material S8).

The TPR and FPR are also calculated theoretically, assuming an underlying AF distribution obtained empirically for each population, as shown in the inset (see Methods for details). The results are shown in Figure 4a. It is seen that the TPR and FPR improve with increasing number of samples albeit with diminishing returns for larger sample sizes. For instance, by constructing a graph using approximately 700 African samples, one can expect to capture 98% of the variation information with a false positive rate slightly above 2%. The experiment and theory agree well, with some deviation for large number of samples due to the limited number of samples available in the experiment. It is important to note that the rate of improvement with increasing number of samples for each population is inversely correlated with the nucleotide diversity of the population (Figure 3). The descending order of nucleotide diversity is $AFR > AMR > SAS > EUR > EAS$, which is exactly the same order of improvement rates, i.e. convergence rates, from low to high. For instance, we observe a slower convergence for the African population as opposed to the European population because the former is much more diverse. The East Asian population graph, on the other hand, improves the fastest since it has the lowest nucleotide diversity. It is also noteworthy that the convergence rate does not depend on the absolute divergence from GRCh38, as mentioned earlier. This is illustrated by the East Asian TPR and FPR curves, which have the highest convergence rates even though the divergence of the East Asian population is higher than all populations except the African. A detailed table for different values of AF cutoff and TPR is provided for all population along with the corresponding FPR in Supplementary Table S1. This table can be used as a guideline for other populations via a comparison of their nucleotide diversity.

In order to measure the effect of sampling on the graph convergence rates, we compare a *homogeneous* sampling approach to a *clustered* sampling approach. Homogeneous sampling ensures that samples for graph construction are picked from all subpopulations uniformly so that there are no abrupt changes in the genetic architecture of the graph reference as more samples are added. Although this approach is expected to more quickly capture the populations genetic information, it is not always possible due to missing detailed ancestry information or different subpopulations being sequenced in different phases of the

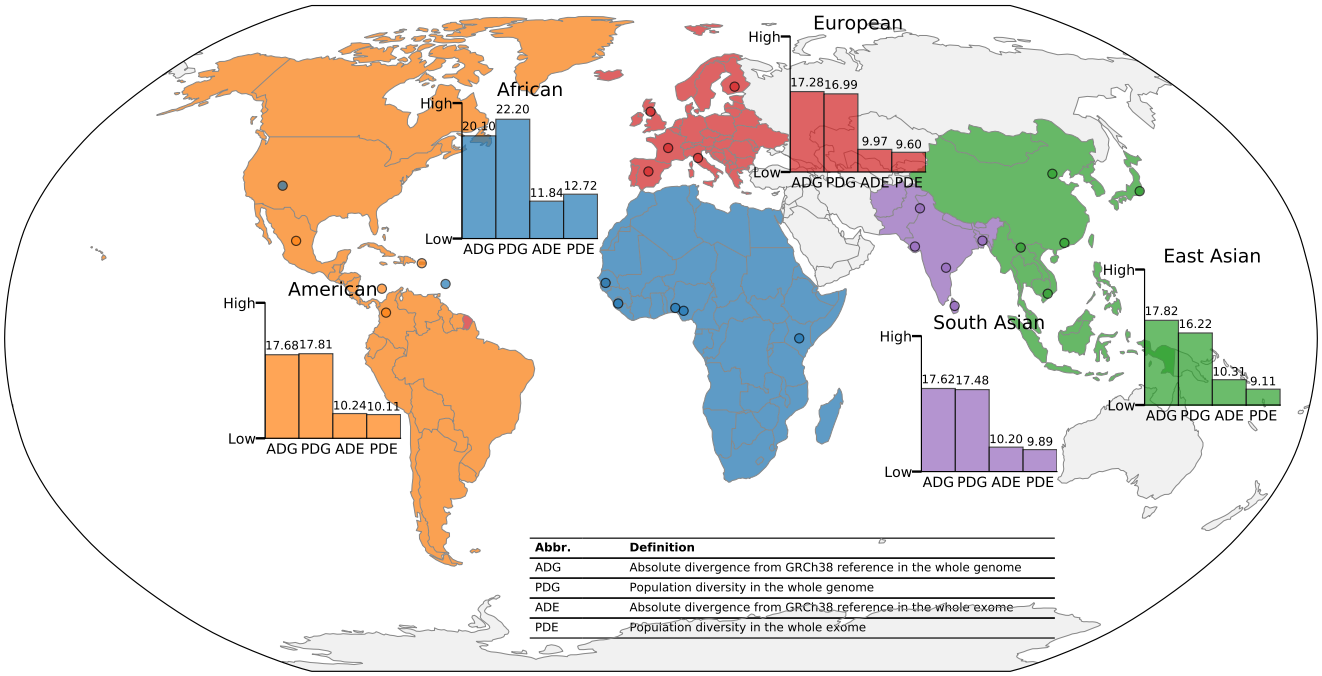

Figure 3. The nucleotide diversity and the absolute divergence from GRCh38 for populations in the 1000 Genomes dataset. Nucleotide diversity measures the average genetic distance between the individuals from the same population and absolute divergence indicates how distant the population is from GRCh38. The African population has the highest genetic diversity and also the largest divergence from the current human genome reference GRCh38, which leads to significant loss of accuracy in pipelines relying on the linear reference. The nucleotide diversity from high to low is as follows: $AFR > AMR > SAS > EUR > EAS$. All populations show sufficiently high diversities and can potentially benefit from a population-specific graph reference.

project. To compare homogeneous sampling to the worst case scenario, clustered approach assumes that the samples are first picked exclusively from a specific subpopulation until there are no samples left belonging to that subpopulation. Then, the procedure is applied to another subpopulation until all subpopulations in the dataset are exhausted. The results are shown in Figure 4b. It is clear that clustered sampling results in a slower convergence rate; therefore, graph references are significantly less representative of the population compared to homogeneous sampling. Abrupt jumps are also observed in TPR and FPR curves at points where a new subpopulation is introduced to the graph reference.

The use of an AF cut-off is a vital filtering step for graph-based methods and justified by the fact that a low frequency variant will pose misinformation to most of the samples and make the alignment more ambiguous and genotyping less accurate for those samples. The exact AF cut-off is a free parameter and can be chosen depending on the specific application, population, or type of sequencing data. We have observed that a value of 5% provides shows good performance without incurring large computational costs (see Supplementary Material S4). For more targeted data such as whole-exome or panel sequencing, a lower threshold can be used without much computational burden.

Graph Reference Construction

Given a set of samples from a population and other relevant variant sources, the method for variant curation is now established as discussed in the previous section. However, the construction of the graph reference itself is not a straightforward task; simply adding all curated variants into the graph structure could cause computational inefficiencies and, at the same time, lead to inaccuracies in alignment and variant calling. In this section, we discuss the pitfalls that can significantly deteriorate the performance of the secondary analysis pipeline and propose a graph construction method that can digest variants of any type (SNPs, MNPs, insertions, deletions, complex structural variants) from multiple sources and produce an optimal graph reference.

The first issue related to incorporating multiple variant databases or VCF files into a graph is variant harmonization. Mainly due to the discordance between the variant representations of different bioinformatics pipelines, the same variants might be expressed differently in different VCF files. Addition of all such variants will introduce different but equivalent paths in the graph reference, leading to alignment problems. To overcome these issues, the multi-allelic variants are split and all variants are left normalized with respect to the backbone of the graph reference (in this case, GRCh38 canonical chromosomes). Moreover,

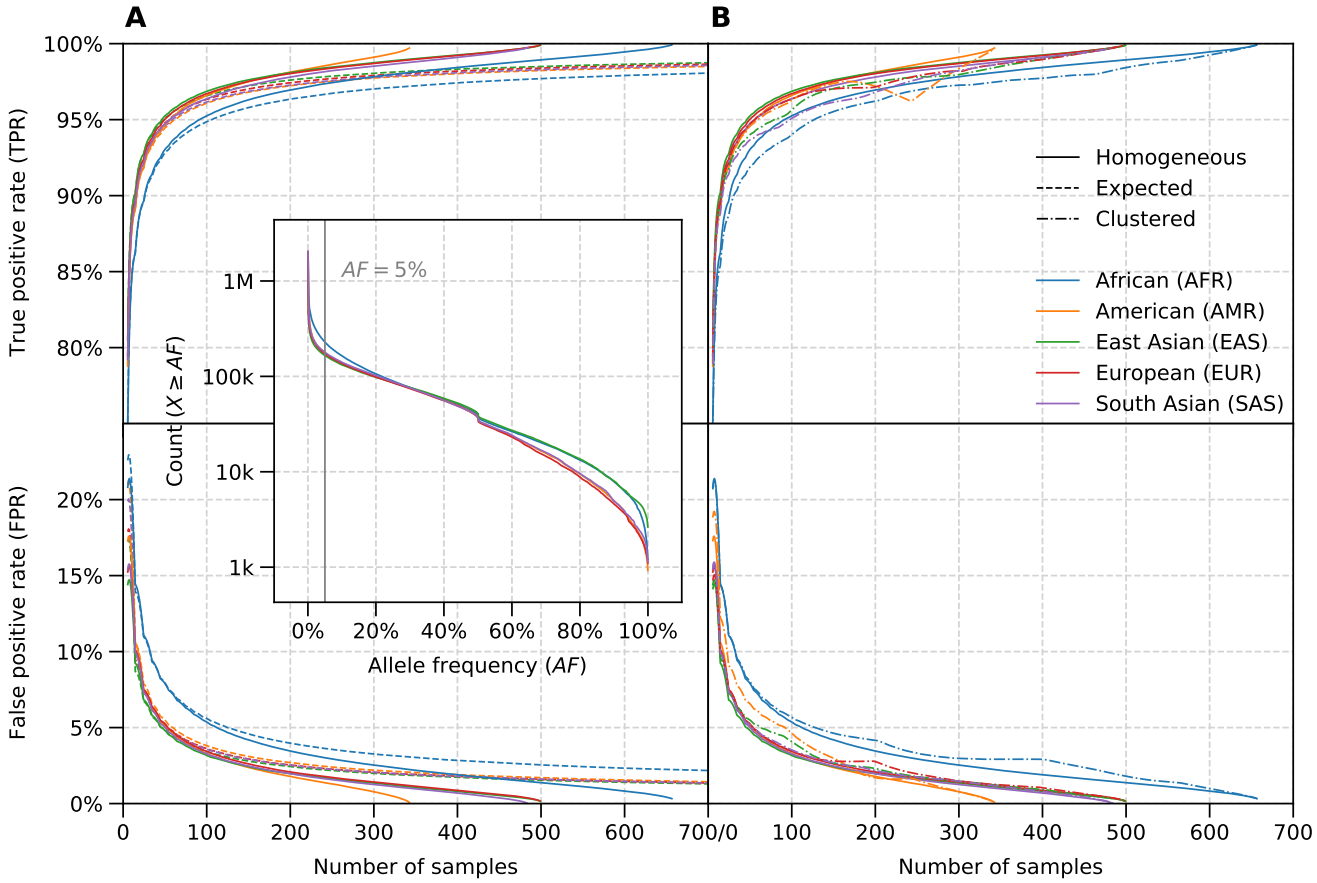

Figure 4. True positive rate (TPR) and false positive rate (FPR) in population-specific graph references with respect to number of samples used for graph construction. TPR and FPR are calculated for each population shown in the legend. An allele frequency (AF) threshold of 5% is used, below which the variants are discarded. (a) Theoretical calculation (see Methods) and simulated results. (b) Simulated results with homogeneous and clustered sampling. Homogeneous sampling assumes samples are taken from subpopulations uniformly, whereas clustered approach simulates the extreme case where samples are taken from each subpopulation until it is exhausted. Inset shows the AF distribution used for calculations, which is obtained empirically from the 1000 Genomes dataset for each major population.

non-standard variant definitions in the VCFs are removed and only fully sequence resolved ones are kept. This process is implemented as the first stage of our graph construction pipeline (*Prepare Inputs*), as shown in Figure 5a.

Another important graph construction step is the preparation of the linear genome reference. It is common to use the linear reference as the backbone of the graph reference, which facilitates variant representation with respect to the same sequence and the coordinate system, ensuring compatibility with the standard bioinformatics tools[29,30,51,52]. We assume the same approach and use the GRCh38 assembly as the backbone for all graphs. There are the so-called alt-contigs in the GRCh38 assembly, which represent alternate sequences for certain regions in the canonical chromosomes. These regions show high variability in the population and alt-contig haplotypes are provided as additional sequences to augment the haploid genome. However, the natural way of incorporating alternate sequences is indeed adding another path to the graph reference. Therefore, we have developed an alt-contig processing step (*GRCh38 Alt Processing* in Figure 5a) which removes alt-contigs from the GRCh38 assembly, maps them to the canonical regions and finally adds them as graph paths with an appropriate representation. The details of this process are shown in Figure 5b. First, the contigs labelled as ALT and NOVEL are removed from the linear reference so that it only contains the primary chromosomes, unplaced and unlocalized contigs and decoy sequences. Next, ALT and NOVEL contigs are mapped to the primary chromosomes. Since they usually contain long stretches of sequences that are identical to the linear reference, ALT and NOVEL contigs are decomposed into smaller variants and left normalized. The final outputs are a modified linear reference that does not contain the alt-contigs and a VCF file that concisely represents alt-contigs with respect to the linear reference.

The *Merge Variants* step in Figure 5a merges the harmonized input variants and the variants obtained from the alt-contigs

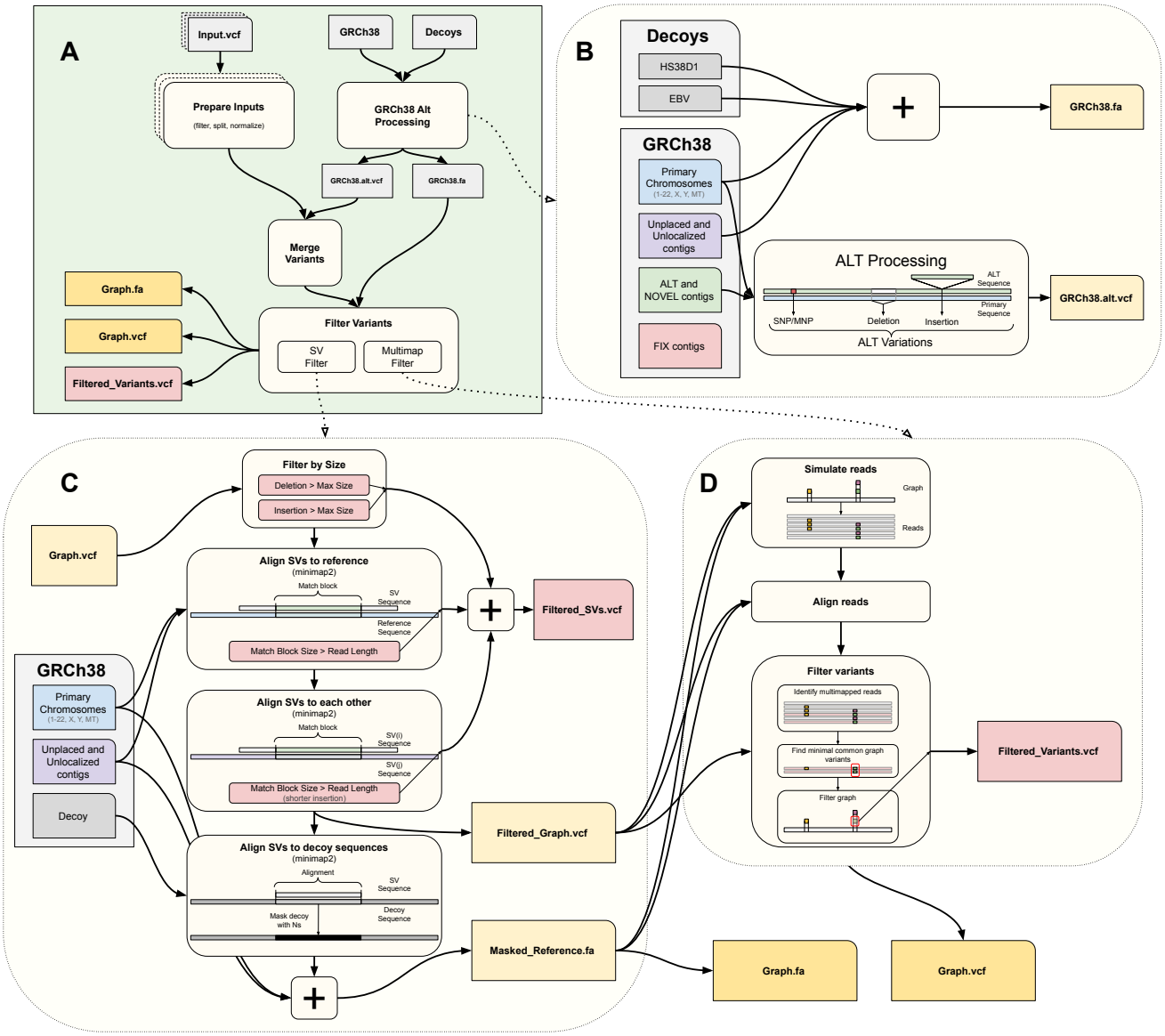

Figure 5. Graph reference construction pipeline. The pipeline takes a set of VCF files and the linear reference (to be used as the backbone of the graph reference) as input, and outputs the constructed graph reference along with the modified linear reference and a list of variants that are excluded from the graph. (i) The input variants are processed to avoid any incompatibility with the graph representation and resolve potential issues in the VCF files. (ii) The alt-contigs in the linear assembly are added as edges into the graph reference and decomposed into smaller variants if necessary. (iii) All variants are merged and the allele frequency is re-calculated. (iv) Structural variants that are similar to the linear assembly or to each other are filtered and the decoy sequences are modified. (v) The reads are simulated from the constructed graph to detect and prune edges that cause multi-mapping in read alignment.

to output a single VCF file containing only biallelic variants without any duplicates. This step also re-calculates the allele frequency for each variant by taking into account all of the input sources and their sample sizes.

After merging all variant sources, the resultant VCF is put through the filtering stage. The first step is the *SV Filter* (see Figure 5a) which processes all structural variants (SVs) to resolve any ambiguity that might be introduced to the reference due to the nature of short-read sequencing data. There are several possibilities to take into account to achieve this: The SVs may be similar to the linear reference; they may be similar to each other; or they may be similar to the decoy sequences which are essentially used to pull all reads that are not supposed to align to the canonical regions. This process is illustrated in Figure 5c. If the match block size between an SV and the reference, which is defined as the longest identical subsequence between the

two, is larger than a threshold, the SV is filtered out, i.e. not added to the graph reference. We choose this threshold to be the read length, since it is very difficult for variant callers to reassemble the reads to detect SVs when there is an alignment gap larger than the read length. Similarly, if the match block size between multiple SVs is larger than the read length, only the largest SV is kept. All SVs that pass these two steps are compared to the decoy sequences in the GRCh38 and any matching sequence is masked in the decoys. This is a necessary step, since the decoy sequences have aggregated not only non-human DNA sequences but also population haplotypes pertaining to under-represented populations such as the African and the Asian ancestries[53]. The SVs can also be filtered due to computational reasons. In this study, we limited the size of insertions and deletions in graphs to 5k and 90k, respectively, for computational efficiency considerations.

A graph reference can be improved by augmenting it with new variant information, as will be shown in the following sections. However, as more variants are added, the number of possible paths in the graph grows exponentially and it becomes highly likely that there will be identical paths in different regions of the graph. This issue introduces ambiguity to the genome during read alignment, potentially causing sequencing reads to be multi-mapped and become uninformative for variant calling. Therefore, a *Multimap Filter* is implemented as the final stage of graph construction as shown in Figure 5d. The purpose of this filter is to break the identity of the paths in the graph by selectively removing variants from it. This is achieved by simulating reads that traverse all possible paths in the graph reference, mapping these reads back to the graph reference, identifying regions that cause multi-mapping, and diluting the variants in these regions of the graph. We calculate the smallest set of variants that resolves the ambiguity and only remove those variants to avoid detracting from the representatives of the graph reference. The exact criteria used in this process is described in Supplementary Section S1, along with the details of all steps shown in Figure 5a.

We use the aforementioned method to construct graph references at each step of the project workflow shown in Figure 2, and look at their contents and growth rates with each iteration. Starting with the public datasets and iteratively augmenting it with the genetic information of the African population using the construction sets, we obtain six different graph references. *Pan-African 0* refers to the population-specific graph obtained using only gnomAD and *Pan-African 5* is the final graph obtained after all five construction sets have been added to the graph. *Pan-African 1* is the first graph that contains variants directly obtained from the cohort under study. In the construction of *Pan-African 1* graph, we also incorporate the high quality SVs curated by the Human Genome Structural Variation Consortium (HGSVC) using PacBio HiFi sequencing data for 10 African samples in the 1000 Genomes dataset[54]. This will further demonstrate in the next section that population-specific graphs can effectively be used for SV genotyping at the population scale.

The variants counts of all graphs, their overlaps with each other and the average AF are shown in Figure 6. Total sizes of each graph are shown on horizontal bars and the intersection sizes are shown for various combinations indicated with filled circles. It is seen that each construction set increases the size of the graph by a significant percentage; however, the growth rate decreases with each iteration. This implies diminishing returns for graph augmentation as more samples are added. Detailed breakdown of each graph's content into variant types can be found in the Supplementary Table S2. The contribution of each source to the final population-specific graph *Pan-African 5* is given in Supplementary Table S3.

The top panel in Figure 6 shows the mean allele frequency (AF) for the corresponding intersection of graphs. The mean AF for the variants added with each subsequent graph augmentation is lower than the previous ones (c.f. the six bars on the right-hand side). This is an expected trend since most of the high frequency variants are already captured in the initial iterations and only the lower frequency variants (above the AF cutoff) remain to be discovered in the rest of the construction set. It should be noted that this trend is observed because the samples are collected uniformly across all subpopulations. If, for instance, two genetically distinct subpopulations of the African population were to be added to the graph reference exclusively at different iterations, then the contents of the graph might change dramatically at an intermediate step, as discussed in the previous section (see Figure 4 for homogeneous and clustered sampling approaches).

Interestingly, each graph contains a small number of variants that are unique to itself. For example, *Pan-African 0* graph contains around 23k variants that are not in any subsequent graphs. This indicates that these variants are removed in the construction of *Pan-African 1* graph because their combination with the new variants added from the first construction set causes ambiguity in the reference.

Finally, we look at the variant composition of the construction sets that are used to construct the graphs *Pan-African 1-5*. The comparison of variants found in each construction set is shown in Figure 7. It is seen that most of the common variants (around 94%) with $AF \geq 5\%$ are shared in all sets as expected, since all sets contain the same number of samples from each African subpopulation and the same female/male ratio. This implies that the processing order of these sets in the project workflow (Figure 2) is unimportant. We also observe a significant reduction in the number of variants shared by subsequent groups and the mean AF of the variants added. This shows that most of the more common variations are captured in earlier iterations, and subsequent groups contribute mostly infrequent variants around AF cut-off.

The total number of variants for each set is shown in the horizontal bars. A consistent increase is observed with each iteration. This indicates that the variant calling sensitivity increases with each graph augmentation and there is merit in

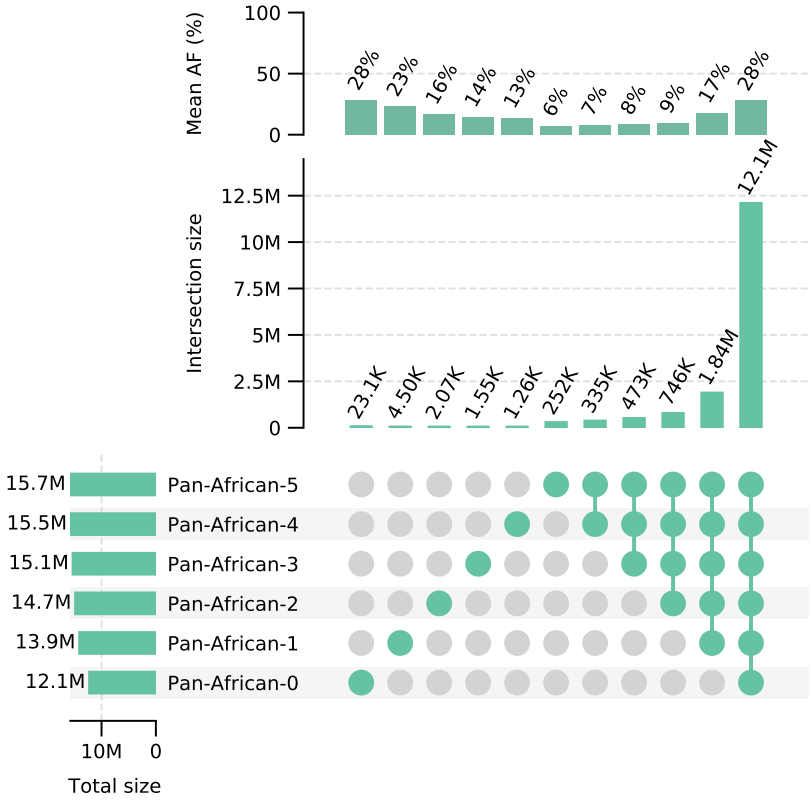

Figure 6. Comparison of constructed graphs. Graphs constructed throughout the iterative process are compared to each other. Total number of variants in graphs are show as horizontal bars on the left. Shared and unique variants in each graph are shown as intersection sets with the number of variants and the mean AF values for each set displayed above.

augmenting graphs even during graph constructions.

Alignment

In order to measure the performance of the constructed graphs on secondary analysis, we use each of them independently to process the benchmarking set of 141 samples, which were left out of graph construction for testing purposes. Specifically, the population-specific graph references *Pan-African 0-5* are compared with each other, with the Pan-Genome graph and also with the linear BWA+GATK pipeline.

First, the alignment accuracy for each pipeline is compared as shown in Figure 8. Each panel shows a different alignment statistic as a violin plot. Each violin corresponds to a different graph reference, representing the median and the distribution of the statistic over all benchmarking samples. Panel (a) shows the percentage of unmapped reads. BWA maps more reads compared to any of the graph references. This is due to the lenient alignment approach used by BWA as opposed to the more stringent criteria used by the graph aligner. All population-specific graphs map more reads compared to the Pan-Genome approach, while progressively mapping more reads with each augmentation. It is seen that improper read (classified as either an improper orientation for read pairs or an insertion length outside the expected range) and uninformative read (MAPQ < 20) percentages are much lower for graph approaches compared to BWA. Population-specific graphs also provide better performance compared to the Pan-Genome graph with the exception of *Pan-African 0*. *Pan-African 0* graph is, although still tailored to the African population, based on public databases and potentially contains many variants irrelevant for the cohort under study. This manifests itself as a larger number of unmapped, improper, and uninformative reads compared to graphs incorporating variants directly obtained from the cohort (*Pan-African 1-5*).

The multi-mapped read ratio is also higher for BWA compared to any graph approach. A distinct jump is observed between iterations 0 and 1. This is due to the addition of cohort-specific SVs into the graph reference. Even though we have implemented a multi-mapping detection filter into the graph construction method (Figure 5d), a graph-based approach will inevitably increase the similarity between genomic regions simply because the genome reference now contains more nucleotide sequences. Our graph construction method effectively balances the trade-off between multi-mapping and other improvements as evidenced by

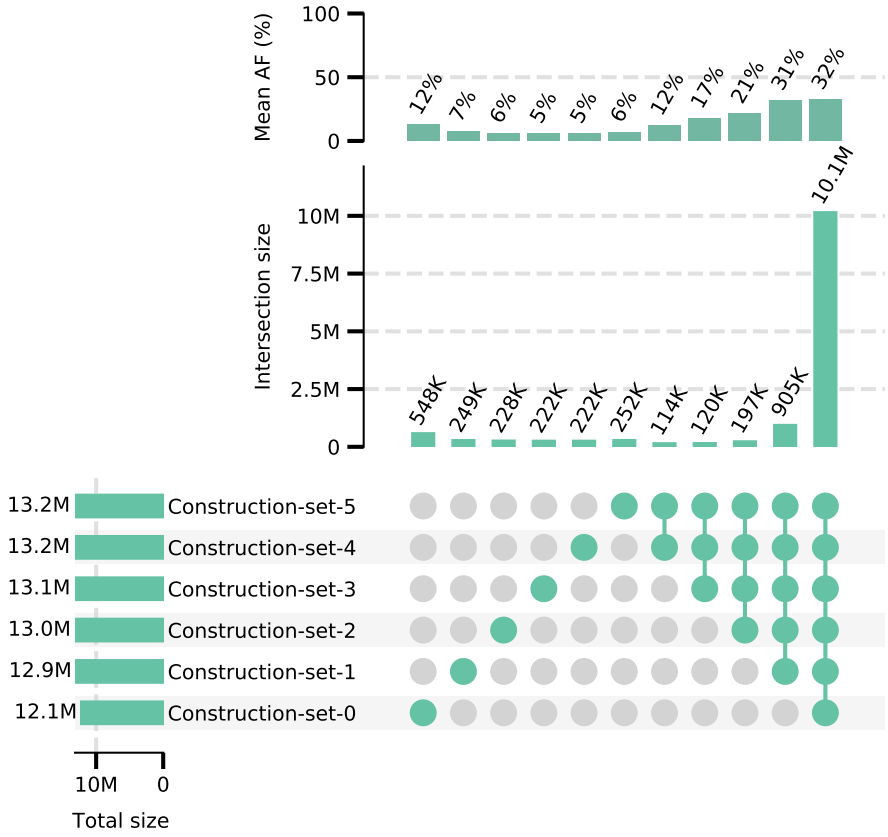

Figure 7. Comparison of construction sets used for graph construction. Each construction set contains common variants (AF$\geq$ 5%) that are selected for addition into the graph. Total number of variants above the AF cutoff in each set is shown as horizontal bars on the left. Shared and unique variants in each set are shown as intersection sets with the number of variants and the mean AF values displayed above.

the significant reduction in the improper and uninformative read ratios from *Pan-African 0* to *Pan-African 1* (Figure 8b & 8d).

Informative reads are defined as any read with a mapping quality larger than or equal to 20 (MAPQ $\geq$ 20). This is a common threshold used in most state-of-the-art variant callers, below which the reads are simply discarded and therefore do not have any influence on variant calling[30,39,55]. It is seen in Figure 8e that graph approaches provide a significantly higher number of informative reads. The difference in the number of informative reads between the graph types is minimal. However, it is equally important to have the reads align to their proper places in addition to having high mapping quality. In this case, the reads are aligning to the population-specific haplotypes in the graphs and there is significant relocation of reads both within chromosomes and between chromosomes (see Supplementary Section S2 for details), which is demonstrated in the next section by the increased sensitivity in variant discovery and SV genotyping.

A useful metric to measure the representativeness of a population-specific is the alignment error rate, i.e. per-base mismatch rate with respect to the genome reference. A smaller error rate indicates that the genetic composition of the population is more successfully captured and also the reference bias is reduced. Figure 8f shows that the error rate consistently decreases from the linear approach to the Pan-Genome graph and to the Pan-African graphs. Each augmentation of the Pan-African graph achieves a better the error rate, leading to around 50% reduction compared to BWA in the last iteration. This is an indication of the accuracy improvements that the iterative graph construction approach can provide.

Next, we investigate how much each graph is utilized during read alignment. We calculate the average number of graph edges that are used in alignment per sample and compare it to the total number of edges in the constructed graph reference, as shown in Figure 9a. The ratio of the two is also shown as the orange line. It is seen that all population-specific graphs have a much higher utilization rate compared to the pan-genome graph which contains variants from many populations. Population-specific graphs provide a utilization rate of around 50% per sample, whereas it is below 40% for the Pan-Genome graph. Notably, even though they are smaller in size, Pan-African 0 & 1 graphs provide better performance than the Pan-Genome graph (see Figures 8 & 10. This implies that a targeted graph reference performs better than a more comprehensive but generic

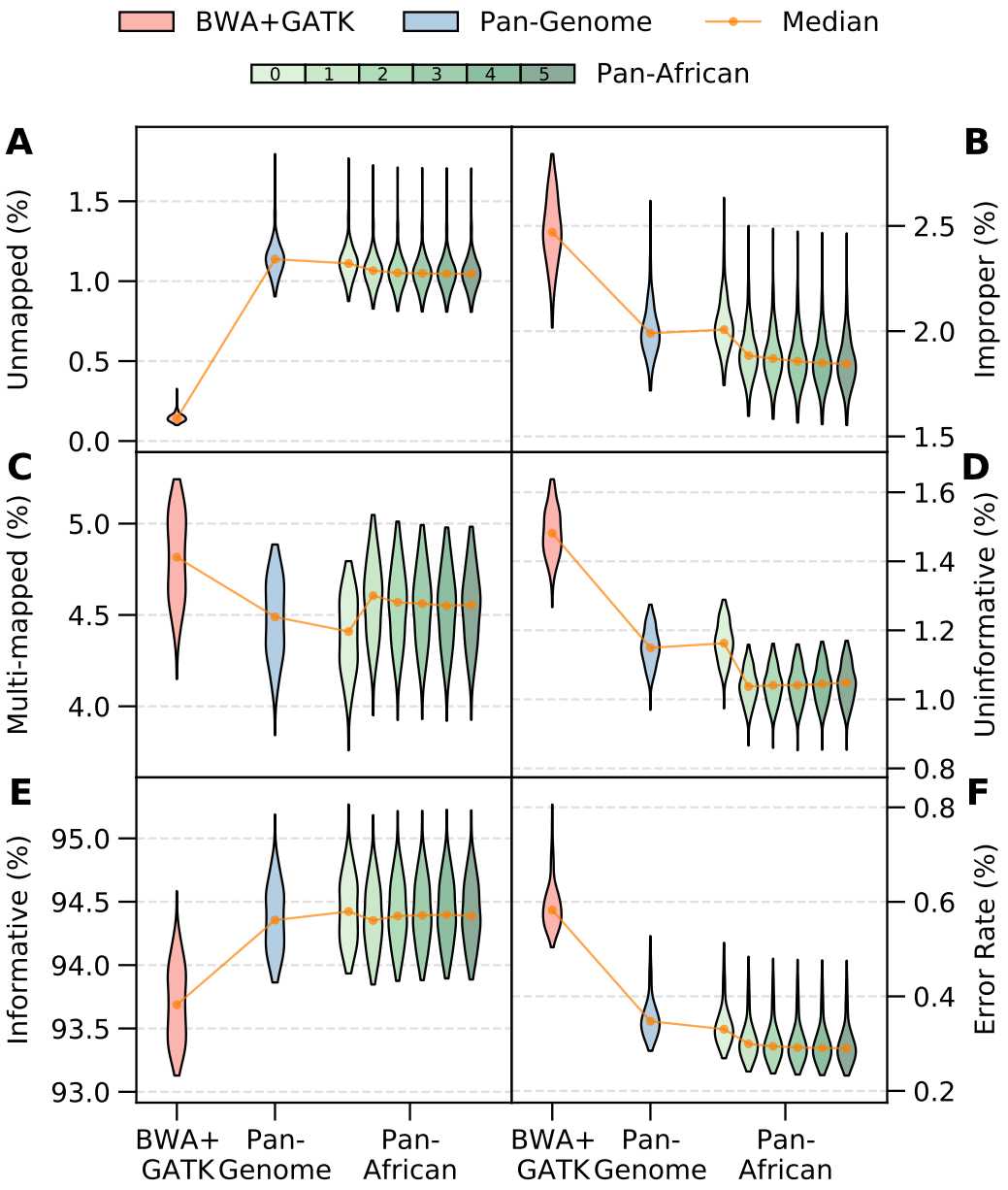

Figure 8. Alignment Metrics. Rate of unmapped (a), improper (b), multi-mapped (MAPQ=0) (c), uninformative (MAPQ < 20) (d) and informative reads (MAPQ ≥ 20) (e). (f) Alignment error rate. Error rate is the ratio of mismatches to aligned bases in alignments with respect to the reference. Wilcoxon tests between consecutive distributions are performed. In all cases except for one (uninformative reads between iterations 2 and 3) the difference is significant ($p < 10^{-3}$).

graph reference, mainly because the irrelevant variants in the generic graph can act as misinformation and cause ambiguity for read alignment. It is also observed that each graph augmentation grows the size of the graph and the number of edges being used, as expected. The utilization rate (orange line) is slowly reduced with each augmentation. This agrees with the results in Figure 6, which shows that the mean AF of the variants added in each subsequent iteration is lower.

Figure 9b shows the total utilization of the graph references across all benchmarking samples. All edges are binned by the number of samples that make use of them in alignment, with yellow and purple bars showing the number of edges used by all and none of the samples, respectively. The Pan-Genome graph has the highest number of unused edges and also the least number of edges used by all samples. The *Pan-African 0* graph contains fewer variants but better utilization, proving that even a population-specific graph based on public datasets constitutes a more representative graph reference. The usage of graph

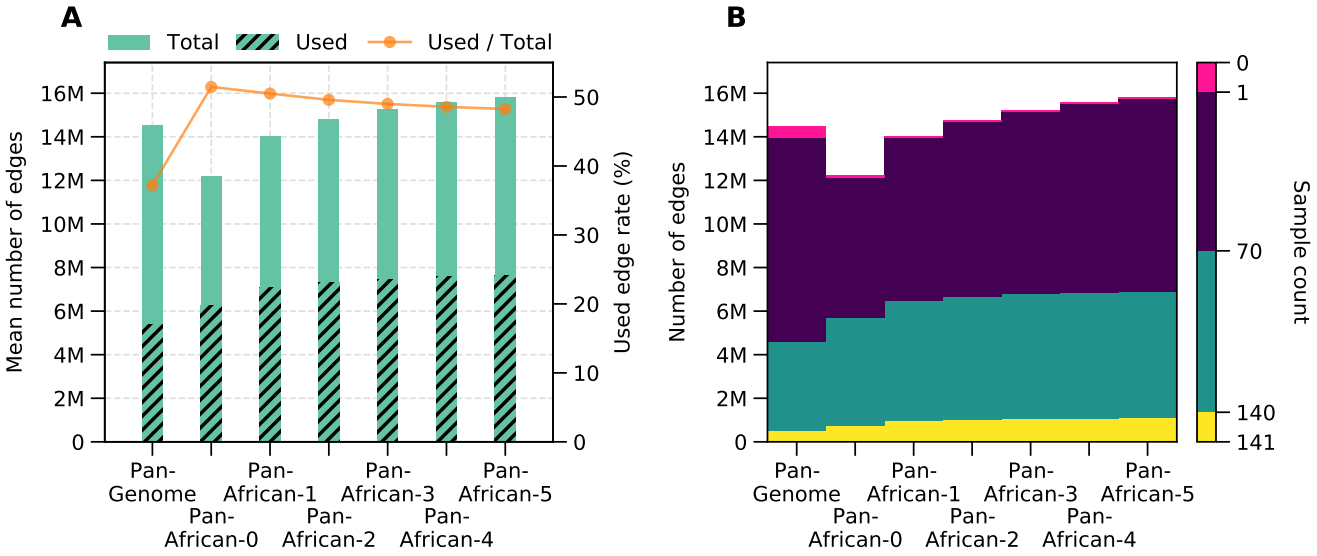

Figure 9. Graph edge utilization. Utilization of graph variants in alignments for benchmark samples. (a) Total number of variants in graph (solid bars) and mean of used variants in alignments per sample (dashed bars). Orange line shows the ratio of used variants with respect to the graph size. (b) Distribution of variant utilization in alignments with respect to the number of samples

edges increase with each augmentation with a trivial amount of unused edges.

Variant calling

In this section, we measure the utility of population-specific graphs for variant calling. In addition to being able to make alignments against graph references, Seven Bridges GRAF pipeline also uses a graph-aware variant caller that can make use of the information stored in graph references. We show the overall performance on single nucleotide polymorphisms (SNPs), insertions and deletions (INDELs) and structural variants (SVs) for all graph references in Figure 10. Panels (a) and (c) show the number of SNPs and INDELs discovered per sample, respectively. It is seen that the Pan-Genome graph provides a higher sensitivity compared to the BWA+GATK pipeline. Moreover, the Pan-African graphs 0 to 5 can increasingly detect more variants compared to the Pan-Genome graph. The Ti/Tv ratio for SNPs is in the expected range and INDEL het/hom ratio approaches the value 3 for the final Pan-African graph. Both SNPs and INDELs are annotated using the dbSNP154 variant database[56] and labelled as known or novel. Fraction of known variants is around 87% for all pipelines, which implies that most of the additional variants detected by the graph pipelines are previously discovered in other studies and therefore are likely to constitute a reliable variant call set. Detailed variant counts categorized into genotypes, variant type and known/novel with respect to dbSNP154 are provided in Supplementary Table S4. The Ti/Tv and Het/Hom ratios for known and novel variants and each variant type are provided in Supplementary Table S10. It is important to note that the values shown in Figure 10B&D are not subject to any additional filtration other than those applied internally by the variant caller. To see the influence of further filtration based on variant call confidence, we filtered the Pan-African-5 graph variant calls based on the QUAL annotation and, in effect, removed lower confidence variants. With a QUAL threshold of 70, the total number of variants per sample is decreased by 116k on average, resulting in a Ti/Tv ratio of 1.95 and an INDEL het/hom ratio of 2.86 (detailed results using various threshold values are provided in Supplementary Table S12).

Figure 10e shows the number of SVs detected by each pipeline (SVs are defined as variants longer than 50 base pairs). The size distribution of SVs are also shown in Figure 10f for BWA+GATK, Pan-Genome, Pan-African 0 and Pan-African 5 pipelines. It is seen that the linear approach BWA+GATK has a significantly lower SV detection rate and can only detect short SVs. The Pan-Genome graph provides considerable improvement over the linear approach. This is made possible by the addition of the alt-contigs in the GRCh38 assembly as alternate paths into the graph reference. Tailoring the graph reference to the African population using public databases that contain only short variants (Pan-African 0) provides only a slight improvement over the Pan-Genome graph and only for short SVs. The major contribution to SV detection in Pan-African 0 graph is still facilitated by the alt-contigs. As discussed previously, starting with Pan-African 1 graph, SVs obtained directly from the 1000 Genomes African samples are used to augment the population-specific graphs. The result of this augmentation is a much higher rate of SV detection and also an increase in the size of SVs that can be genotyped.

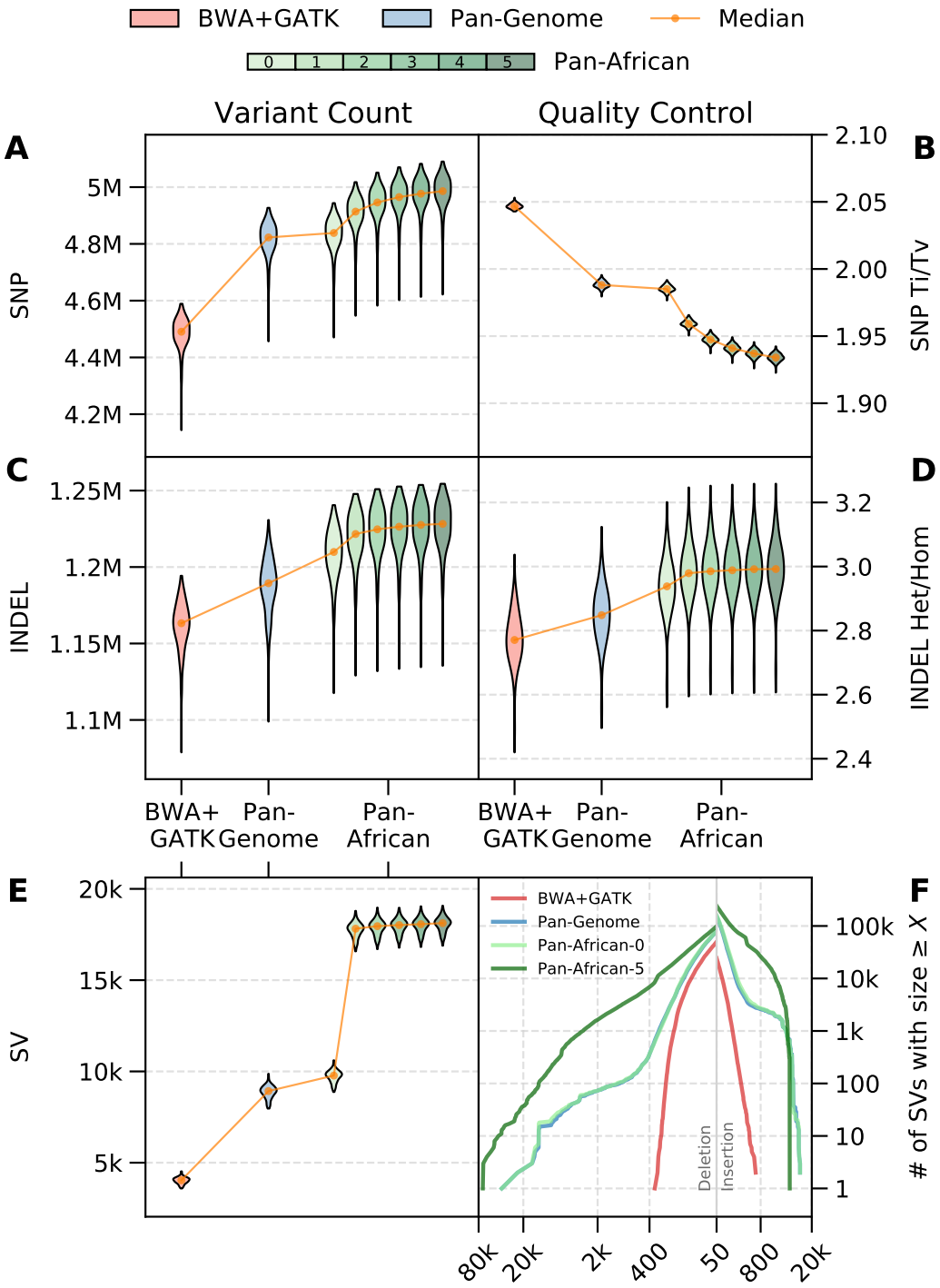

Figure 10. Variant calling metrics. Sample distribution of SNP counts (a), transition to transversion ratio (Ti/Tv) for SNPs (b), INDEL counts (c), heterozygous to non-reference homozygous ratio (Het/Hom) for INDELs (d) and structural variant (SV) counts (e). Panel (f) shows the size distribution of detected structural variants. Wilcoxon tests between consecutive distributions are performed. In all cases the difference is significant ($p < 10^{-21}$).

Based on the results comparing different versions of population-specific graphs, it is concluded that the each iteration provides better overall performance in terms of alignment and variant calling accuracy. Using the output of the last iteration as the final graph reference, we now compare the variant calls made by the *Pan-African 5* and the BWA+GATK pipelines in more

detail. Figure 11 shows the cumulative variant counts for both pipelines with respect to the allele frequency (detailed counts are

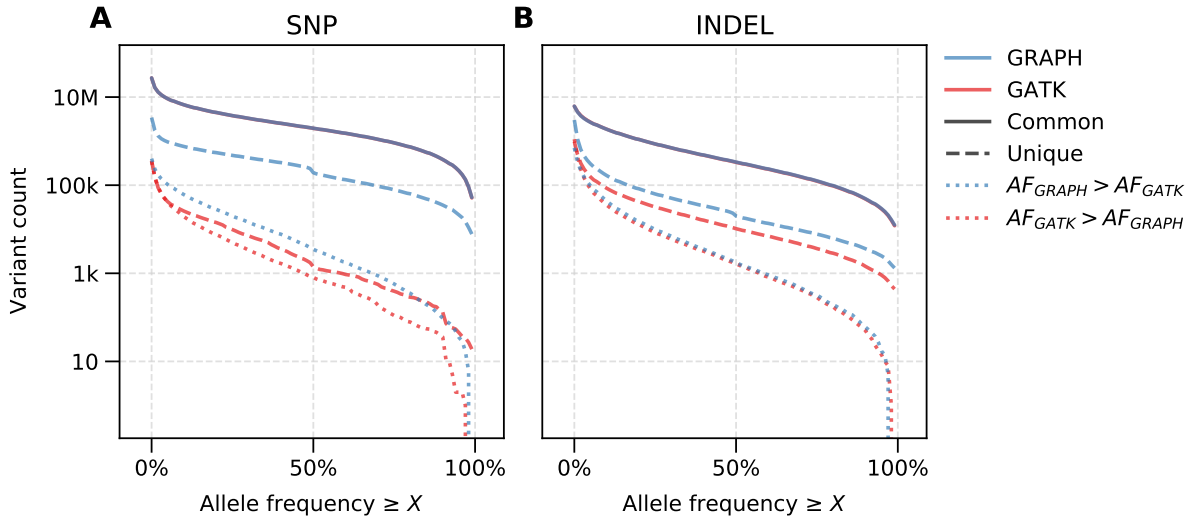

Figure 11. Comparison of variant calls for Graph (Pan-African 5) and GATK results. Cumulative variant counts with respect to the allele frequency are shown for the graph (blue) and the GATK (red) pipelines. Solid lines show the common calls between the two pipelines. Dashed lines show the unique variants to each pipeline. Dotted lines indicate the common variants that are called with a different AF by the two pipelines as shown in the legend, i.e. the blue (red) dotted line shows the variants for which graph (gatk) detects a higher allele frequency.

provided in Supplementary Table S7). The variants are first classified into SNPs and INDELs (panels **A** and **B**, respectively), and then into common (detected in the population by both pipelines) and unique (detected by either pipeline) variant sets. High concordance is observed between the pipelines as majority of the variants are detected by both pipelines (solid lines). In order to distinguish between the genotyping efficacy of these methods, common variants are further split into two categories as $AF_{GRAF} > AF_{GATK}$ and $AF_{GATK} > AF_{GRAF}$ (dotted lines). The former represents the number of variants that are detected in the population by both methods but genotyped with a higher sensitivity by the graph pipeline (and vice versa for the latter). Among the variants observed in the population with a high frequency ($\geq 5\%$), graph pipeline is able to genotype approximately 120k INDELs and 119k SNPs with a higher AF, where as the same numbers for GATK is 106k INDELs and 51k SNPs. Additionally, it is noteworthy that the graph-based approach identifies approximately 6 times as many unique variants as the linear method.

In order to predict the potential clinical significance of the variants detected by the graph-based approach and rule out any bias in variant calling sensitivity towards specific genomic regions or prevalence in population, we stratify all detected variants detected into exonic, intronic and intergenic regions. We further divide the variants into three frequency bins as singleton (observed in only one sample), rare ($AF < 5\%$ but observed in multiple samples) and common ($AF \geq 5\%$), and compare the results to the linear approach BWA+GATK. Table 1 shows the counts of variants unique to each pipeline (detailed variant counts are provided in Supplementary Table S5).

The use of Pan-African graph leads to the detection of 3 to 4 times more high and moderate impact variants in exonic regions for all frequency bins, compared to the BWA+GATK pipeline. Specifically, there are 429 and 9457 more high and moderate impact variants, respectively, detected by the graph pipeline. The contrast between the linear and the graph approaches increases in intronic and intergenic regions although the total number of high and moderate impact variants is lower in these regions. Similar trends are observed for low impact and modifier mutation events in all regions. These observations indicate that the Pan-African graph improves the sensitivity across the whole genome without any bias towards a specific genomic region.

Graph-based analysis as an alternative to joint calling

The standard practice for analyzing the sequencing data of a large pool of whole-genome or whole-exome samples is using the variation information at the population level to inform genotyping at the individual level. This is commonly achieved by using *joint calling*: a method that produces a genomic VCF (GVCF) file from read alignments for each sample independently, and then jointly genotypes all GVCF files together[39]. The result is a multi-sample VCF file containing genotype information for all samples at each loci where there is variation in the population. One of the fundamental advantages of joint calling is that it can recover missing variants and correct genotypes in individuals by considering the rest of the population. The recommended step after joint calling is the *variant quality score recalibration (VQSR)*: a machine learning based variant filtration tool that is

Region	Impact	Singleton (Graph)	Rare (Graph)	Common (Graph)	Singleton (GATK)	Rare (GATK)	Common (GATK)
Exonic	HIGH	935	259	145	419	294	197
Exonic	MODERATE	5855	3582	2944	1371	1028	525
Exonic	LOW	2533	1698	1497	570	404	270
Exonic	MODIFIER	36781	29140	20203	5354	6949	2841
Intronic	HIGH	297	218	158	39	31	17
Intronic	LOW	1310	982	730	159	171	133
Intronic	MODIFIER	1239286	998546	404136	255666	321315	92814
Intergenic	MODERATE	458	389	206	100	102	17
Intergenic	MODIFIER	1575710	1287305	831713	240360	300409	99425

Table 1. Functional impact of detected variants that are unique to each pipeline. Variants are split based on their occurence; Singleton (observed in a single sample), Rare (AF $< 5\%$) and Common (AF $\geq 5\%$).

effective on large number of samples[39].

A population-specific graph reference readily captures the population's genetic diversity and therefore it is able to utilize this information not just in variant calling but also in read alignment, unlike the standard joint calling approach. To measure how these two approaches compare to each other, we extract the variants that are recovered/corrected by the BWA+GATK joint calling pipeline, i.e. variants that would have been missed with single sample calling, and look at the concordance of these variant calls with the calls made by the graph-based approach. The variants are classified by VQSR as PASS and non-PASS,

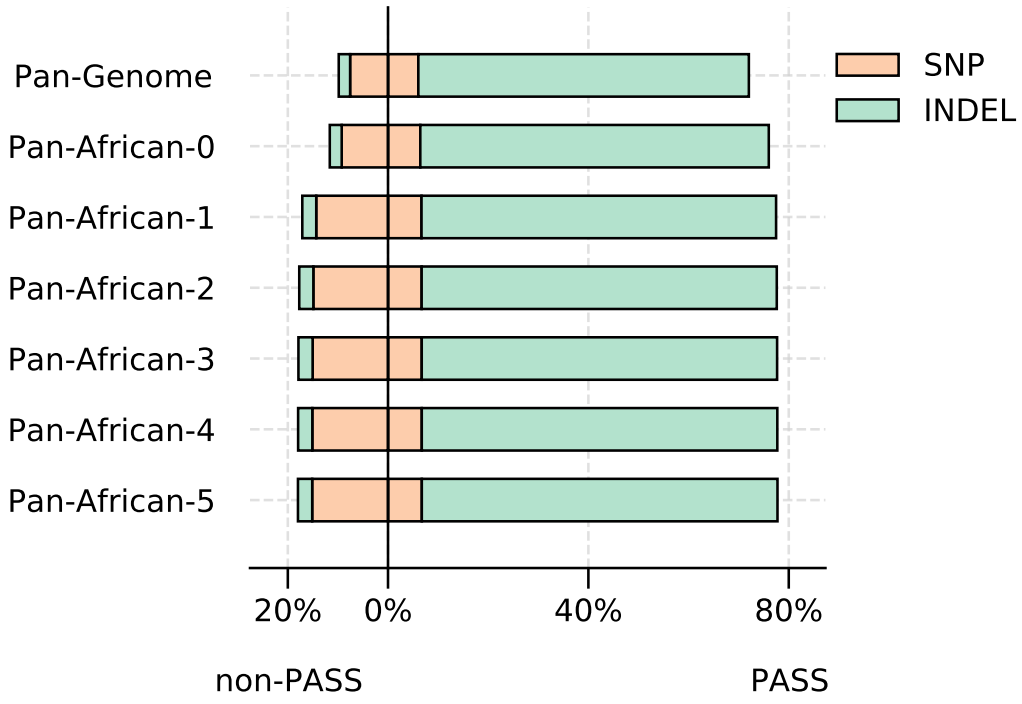

Figure 12. Concordance between GATK joint calling and population-specific graph approach. Percent of loci called by the graph pipeline for the variants rescued in traditional joint calling. Results are split based on the filtration output of VQSR.

indicating whether or not they pass the filtration criteria to be considered as high quality variant calls. The percentage overlap of these recovered/corrected genomic loci with the calls made by each graph pipeline is shown in Figure 12. The Pan-African 5 graph is able to genotype almost 80% of the variants recovered by traditional joint calling, without the need of a post-processing step, while calling less than 20% of the variants filtered out by VQSR. Since a population-specific graph contains variants relevant only to the population under study, it can provide both sensitivity and specificity. All population-specific graphs after the second iteration show similar concordance with respect to joint calling. It is also noteworthy that even the Pan-Genome

graph provides most of the improvements provided by joint calling. The detailed breakdown of variant counts for each genotype and variant type is provided in Supplementary Table S6.

Discussion

The conceptual distinction between the nucleotide diversity within a population and the absolute divergence from a reference sequence plays a decisive role in setting the expectations for the accuracy of sequencing data analysis and approaches that utilize novel reference structures such as genome graphs. Despite the quantitative similarity, these two metrics are fundamentally independent. Nucleotide diversity measures the average genetic distance between any pair of individuals from the same population and therefore carries a biological meaning. In the context of NGS, the reference genome is used only to have an intermediate representation of an individual's genome, therefore it does not influence, at least theoretically, the diversity measurement. On the other hand, absolute divergence is the genetic distance to an arbitrarily defined DNA sequence (GRCh38 in this case), and its value can range from very low to very high for populations with a similar diversity, as empirically shown in Figure 3. Although it does not have a biological foundation, the absolute divergence significantly influences the performance of bioinformatics methods. Standard approaches can suffer from a loss of accuracy when used on divergent populations, which is usually reflected in increased reference bias in alignment and decreased variant call sensitivity. The variant curation method for graph construction presented in this study lays out a procedure for sample selection that is based on the population's genetic diversity. The representativeness of the resultant genome graph is also calculated with associated true and false positive rates. We expect that the improvements obtained from a genome graph will be more dramatic for divergent populations. We have tested the graph construction method on the African population which is both the most diverse and the most divergent among the five populations in the 1000 Genome dataset. While exemplifying the suitability of genome graphs for large scale sequencing projects, the iterative graph construction approach emphasizes the importance of extracting genetic information directly from the cohort under study and making the graph reference more tailored to the cohort. Transitioning from the linear human genome reference GRCh38, which is the least specific reference, to a pan-genome and finally to a set of population-specific graphs is shown to improve secondary analysis on multiple fronts. We expect this improvement trend to continue as graphs are further tailored to sub-populations or even smaller groups of individuals with higher genetic similarity, assuming the number of samples in the group remains sufficiently large to capture the genetic context.

There are several pitfalls faced during the construction of a graph reference. With an aim to address these, our graph construction method takes into account the possibility of introducing ambiguity to the reference genome as more variants are added and resolves the culpable graph paths. This is a crucial step regardless of the types of variants being added to the graph; we have observed in the construction of Pan-African graphs that even a few SNPs might cause undesirable amounts of read multi-mapping. Moreover, great care should be exercised if large variants such as SVs are being added to the graph to make sure that they do not have high similarity to the reference genome or to each other. These control mechanisms lead to an effective graph construction with alignment benefits such as higher rate of informative reads and reduced reference bias, which remains an obstacle for linear methods, especially in genomic studies of under-represented populations. Since the applicability of our graph construction method (see Figure 5) is not limited to the specific tools used in this study, we expect that similar improvements can be obtained for other graph-based bioinformatics toolkits.

We have shown that population-specific genome graphs facilitate the detection of more variants and genotyping of SVs at the population scale. We have been able to identify thousands of functionally important variants in the exome that are completely missed by the standard BWA+GATK pipeline. Additionally, we have detected significantly more novel SNPs and INDELs when compared against dbSNP database release v154 (see Supplementary Table S4). Further studies are required to truly understand the impact of these variants for clinical applications. Another advantage of population-specific graphs is that they can readily provide the sensitivity and specificity improvements expected from joint calling without the need for simultaneous processing of all samples in the cohort. This also removes the computational burden faced by joint calling when applied to large cohorts on the order of tens of thousands of whole genome samples.

Methods

Nucleotide Diversity and Absolute Divergence

Graphs rely on enhancing linear reference with common polymorphisms existing within a population. Therefore it is important to estimate the level of polymorphism existing within a population (*diversity*) as well as how much the population differs from the linear reference (*divergence*).

Polymorphism within a population is commonly measured with *nucleotide diversity* defined by Nei and Li in 1979[47]. This measure calculates the average number of nucleotide differences between two sequences for all possible pairs within a population. Since variant call format defines sequence differences with respect to a common linear reference, nucleotide diversity can simply be calculated from variant calls of samples from a population. For a given variant locus, diversity

contribution will be sum of the number of base differences between two alleles weighted with respect to their occurrence frequencies over all possible allele pairs at that loci. Diversity for a genomic region will be the sum of all diversity contributions from variant loci in the region divided by the size of the region.

$$Diversity = \frac{\sum_{\text{variant loci } \in R} \sum_{i,\ j \neq i} \frac{|i||j|\delta_{i,j}}{N(N-1)}}{|R|} \tag{1}$$

where i, j are the distinct alleles at given loci, $|i|$ is the number of occurrences for particular allele, $\delta_{i,j}$ is the edit distance between two alleles, N is the total number of alleles at given loci and $|R|$ is the number of bases within the genomic region.

Divergence is similar to diversity, but instead of measuring the differences between two samples from the population, each sample in the population is compared against the linear reference. Therefore the divergence contribution at a variant locus is average nucleotide differences between an allele and the reference weighted by the occurence frequencies of those alleles. Divergence within a genomic region is then the sum of all divergence contributions divided by the size of the region.

$$Divergence = \frac{\sum_{\text{variant loci}} \sum_{i} \frac{|i|\delta_{i,r}}{N}}{|R|} \tag{2}$$

where $\delta_{i,r}$ is the edit distance between allele i and referece allele at that loci.

Calculation of TPR and FPR in Graph Construction

Graphs enhance linear reference by incorporating common variations, therefore variants are selected by a allele frequency cut-off. Furthermore, since graphs are constructed from a subset of samples from a population, observed allele frequency of a variant in the subset can differ from the ideal allele frequency in whole population thus resulting in a different set of variants to be selected for graph. Variants in the graph constructed from a subset can be divided into two groups: *true* variants (ideal allele frequency above cut-off) and *false* variants (ideal allele frequency below cut-off). True positive rate (TPR) is the rate of *true* variants with respect to ideal graph size, and similarly false positive rate (FPR) is the rate of *false* variants with respect to ideal graph size.

Ideal allele frequency of a given variant can be considered as the occurrence probability of that variant for each allele. Assuming the occurrence of variant is independent for each allele, number of times a variant is observed in N diploid samples ($2N$ alleles) follows a binomial distribution with allele frequency as success probability. For an allele frequency cut-off (f_c) used in graph construction, the probability of adding a variant with true allele frequency f into graph in N samples is then sum of probabilities where occurrence count results in observed allele frequency larger than cut-off ($k/2N \geq f_c$).

$$P(added|f) = P\left(\frac{k}{2N} \geq f_c | f\right) = \sum_{k \geq 2N f_c}^{2N} \binom{2N}{k} f^k (1-f)^{2N-k} \tag{3}$$

TPR can be calculated by the expected fraction of true variants added to graph in N samples divided by the fraction true variants.

$$TPR = \frac{\int_{f_c}^{1} P(added|f)p(f)df}{\int_{f_c}^{1} p(f)df} \tag{4}$$

Similarly, FPR can be calculated by the expected fraction false variants added to graph in N samples divided by the fraction of true variants.

$$FPR = \frac{\int_{0}^{f_c} P(added|f)p(f)df}{\int_{f_c}^{1} p(f)df} \tag{5}$$

Additional information

Supplementary information is available.

Competing financial interests: All authors have been employed by Seven Bridges Inc. throughout the period of work for this study.

References

1. Consortium, I. H. G. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. Venter, J. C. *et al.* The sequence of the human genome. *science* **291**, 1304–1351 (2001).
3. Hinds, D. A. *et al.* Whole-genome patterns of common dna variation in three human populations. *Science* **307**, 1072–1079 (2005).
4. Consortium, I. H. *et al.* A haplotype map of the human genome. *Nature* **437**, 1299 (2005).
5. Fonseca, N. A., Rung, J., Brazma, A. & Marioni, J. C. Tools for mapping high-throughput sequencing data. *Bioinformatics* **28**, 3169–3177 (2012).
6. Lee, I.-H. *et al.* Prioritizing disease-linked variants, genes, and pathways with an interactive whole-genome analysis pipeline. *Human mutation* **35**, 537–547 (2014).
7. Consortium, . G. P. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
8. consortium, U. *et al.* The uk10k project identifies rare variants in health and disease. *Nature* **526**, 82 (2015).
9. Consortium, W. T. C. C. *et al.* Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661 (2007).
10. Nagasaki, M. *et al.* Rare variant discovery by deep whole-genome sequencing of 1,070 japanese individuals. *Nature communications* **6**, 1–13 (2015).
11. Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the icelandic population. *Nature genetics* **47**, 435–444 (2015).
12. Schneider, V. A. *et al.* Evaluation of grch38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome research* **27**, 849–864 (2017).
13. Green, R. E. *et al.* A draft sequence of the neandertal genome. *science* **328**, 710–722 (2010).
14. E pluribus unum. *Nat Methods* **7**, 331 (2010).
15. Ballouz, S., Dobin, A. & Gillis, J. A. Is it time to change the reference genome? *Genome biology* **20**, 1–9 (2019).
16. Bentley, A. R., Callier, S. L. & Rotimi, C. N. Evaluating the promise of inclusion of african ancestry populations in genomics. *NPJ genomic medicine* **5**, 1–9 (2020).
17. Rosenfeld, J. A., Mason, C. E. & Smith, T. M. Limitations of the human reference genome for personalized genomics. *PloS one* **7**, e40294 (2012).
18. Kehr, B. *et al.* Diversity in non-repetitive human sequences not found in the reference genome. *Nature Genetics* **49**, 588–593 (2017).
19. Computational pan-genomics: status, promises and challenges. *Briefings in bioinformatics* **19**, 118–135 (2018).
20. Yang, X., Lee, W.-P., Ye, K. & Lee, C. One reference genome is not enough. *Genome biology* **20**, 104 (2019).
21. Rozowsky, J. *et al.* Alleleseq: analysis of allele-specific expression and binding in a network framework. *Molecular systems biology* **7**, 522 (2011).
22. Vijaya Satya, R., Zavaljevski, N. & Reifman, J. A new strategy to reduce allelic bias in rna-seq readmapping. *Nucleic acids research* **40**, e127–e127 (2012).
23. Huang, L., Popic, V. & Batzoglou, S. Short read alignment with populations of genomes. *Bioinformatics* **29**, i361–i370 (2013).
24. Maretty, L. *et al.* Sequencing and de novo assembly of 150 genomes from denmark as a population reference. *Nature* **548**, 87–91 (2017).
25. Sherman, R. M. *et al.* Assembly of a pan-genome from deep sequencing of 910 humans of african descent. *Nature genetics* **51**, 30–35 (2019).
26. Duan, Z. *et al.* Hupan: a pan-genome analysis pipeline for human genomes. *Genome biology* **20**, 149 (2019).
27. Schneeberger, K. *et al.* Simultaneous alignment of short reads against multiple genomes. *Genome biology* **10**, 1–12 (2009).
28. Paten, B., Novak, A. M., Eizenga, J. M. & Garrison, E. Genome graphs and the evolution of genome inference. *Genome research* **27**, 665–676 (2017).

29. Garrison, E. *et al.* Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature biotechnology* **36**, 875–879 (2018).

30. Rakocevic, G. *et al.* Fast and accurate genomic analyses using genome graphs. *Nature genetics* **51**, 354–362 (2019).

31. Eggertsson, H. P. *et al.* Graphtyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nature communications* **10**, 1–8 (2019).

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

33. 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).

34. Pritt, J., Chen, N.-C. & Langmead, B. Forge: prioritizing variants for graph genomes. *Genome biology* **19**, 1–16 (2018).

35. Gaziano, J. M. *et al.* Million veteran program: A mega-biobank to study genetic influences on health and disease. *Journal of clinical epidemiology* **70**, 214–223 (2016).

36. Bycroft, C. *et al.* The uk biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

37. Hutter, C. & Zenklusen, J. C. The cancer genome atlas: creating lasting value beyond its data. *Cell* **173**, 283–285 (2018).

38. Snyder, M. P. *et al.* Perspectives on encode. *Nature* **583**, 693–698 (2020).

39. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv* 201178 (2017).

40. Li, H. *et al.* The sequence alignment/map format and samtools. *Bioinformatics* **25**, 2078–2079 (2009).

41. Danecek, P. *et al.* The variant call format and vcftools. *Bioinformatics* **27**, 2156–2158 (2011).

42. Amstutz, P. *et al.* Common workflow language, v1. 0 (2016).

43. Birney, E., Vamathevan, J. & Goodhand, P. Genomics in healthcare: Ga4gh looks to 2022. *BioRxiv* 203554 (2017).

44. Li, H. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv preprint arXiv:1303.3997* (2013).

45. Byrska-Bishop, M. *et al.* High coverage whole genome sequencing of the expanded 1000 genomes project cohort including 602 trios. *bioRxiv* 2021.02.06.430068 (2021).

46. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

47. Nei, M. & Li, W.-H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences* **76**, 5269–5273 (1979).

48. Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The oxford nanopore minion: delivery of nanopore sequencing to the genomics community. *Genome biology* **17**, 1–11 (2016).

49. Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature biotechnology* **37**, 1155–1162 (2019).

50. Eid, J. *et al.* Real-time dna sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).

51. Kim, D., Langmead, B. & Salzberg, S. L. Hisat: a fast spliced aligner with low memory requirements. *Nature methods* **12**, 357–360 (2015).

52. 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).

53. Mallick, S. *et al.* The simons genome diversity project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).

54. Ebert, P. *et al.* Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* (2021).

55. Poplin, R. *et al.* A universal snp and small-indel variant caller using deep neural networks. *Nature biotechnology* **36**, 983–987 (2018).

56. Sherry, S. T. *et al.* dbsnp: the ncbi database of genetic variation. *Nucleic acids research* **29**, 308–311 (2001).