Unveiling Genomic Complexity: A Framework for Genome Graph Structural Analysis and Optimised Variant Calling Workflows

Venkatesh Kamaraj[1,2], Ayam Gupta[1,2], Karthik Raman[1,2,3,4*], Manikandan Narayanan[1,2,5*], Himanshu Sinha[1,2,3*]

[1]Centre for Integrative Biology and Systems Medicine (IBSE), [2] Wadhwani School of Data Science and AI, [3] Department of Biotechnology, Bhupat and Jyoti Mehta School of Biosciences, [4] Department of Data Science and AI, [5] Department of Computer Science and Engineering, IIT Madras, Chennai, India

* *Co-corresponding authors: kraman@iitm.ac.in, nmanik@cse.iitm.ac.in, sinha@iitm.ac.in*

ABSTRACT

Genome graphs offer a powerful alternative to linear reference genomes, as they provide a richer representation of a collection of genomes by emphasising the polymorphic regions. Despite their advantages, it is difficult to characterise the structural complexity of a genome graph and efficiently call variants using it. In our study, we developed a novel framework to visualise and quantify the structural complexity of genome graphs and optimised existing genome graph-based variant-calling workflows. We applied these techniques on two human pan-genome graphs constructed with the variants from the 1000 Genomes Project - the first augmenting only common variants and the second all the variants, including rare ones. We found that rare variants increased the variability of a genome graph by 10-fold but hypervariability by 50-fold. Our framework highlighted biologically significant and highly variable loci like HLA and DEFB and several novel variable regions. Benchmarking the optimised variant-calling workflow with the two genome graphs identified 9.83% more variants than the linear reference genome. Additionally, it showed that rare variants had minimal impact on variant-calling performance. Our methods are effective in identifying novel genomic loci and variants from genome graphs and revealing the underlying genomic complexity of the represented population or species.

INTRODUCTION

The reference genome is a cardinal element in genome analysis. Despite its broad impact, the first draft of the reference genome (1), unveiling the 3 billion base pairs of DNA, missed much of the variation unique to heterogeneous human populations (2). Since the release of the first draft, the reference genome has been improved regularly to redress its inadequacies. The current widely used human reference genome GRCh38, released in 2013 (3) and recently updated in 2019 (4), is a haploid DNA sequence majorly derived from merged haplotypes from individuals of Caucasian and African ancestries, with a single individual comprising most of the sequence (5, 6). Various studies (7–10) elucidate the genetic diversity and heterogeneity of human populations. However, recent advances in genomics have shown that owing to its haploid linear structure, the current reference cannot sufficiently capture the genetic diversity in human populations (11). As only one allele of a variant can be present in a haploid genome, using such a reference structure leads to reference allele bias (12). These drawbacks of a linear reference genome highlight the need for a better reference structure to efficiently represent the sequence information in the DNA of heterogeneous and diverse human populations.

Genome graphs have the potential to represent highly heterogeneous populations while overcoming the drawbacks that usually accompany using traditional linear structure-based references (13). Variational genome graphs encode the genetic variants within a population and embed paths that can represent possible sequences from a population for which the reference is built (14, 15). They methodically segment sequences into discrete entities termed nodes. The interconnection of these nodes through edges establishes a network, facilitating seamless traversal within the graph from one node to another. Genome graphs can incorporate information from multiple individuals from varied populations, making them more versatile for studying diverse genomic landscapes (16). Previous studies broadly classify genome graphs as population-specific, capturing the genetic variations like single nucleotide polymorphisms (SNPs), insertions, deletions (INDELs), and structural variants (SVs) that are prevalent within a particular population and pan-genomes representing the genetic diversity across multiple populations (17). As genome graphs represent alternate alleles at specific genomic loci, they overcome reference allele bias (12), enabling superior variant calling and more accurate downstream analysis (18). Yet, existing variant calling workflows that are based on genome graphs are running time intensive.

Genome graphs also make it feasible to focus on hypervariable genomic regions at a population and personalised level. Despite this inherent advantage, methods to comprehensively study the structural complexities underlying genome graphs are lacking. Previous studies have used metrics such as the total number of nodes, edges and connected components as indicators of the complexity of the graph (19). Although helpful in explaining the overall complexity, these metrics do not represent the intricate and localised complexity intrinsic to genome graphs. A comprehensive framework is a prerequisite to elucidate the genetic diversity of the represented species or population. Moreover, existing visualisation techniques are limited to small genome graphs and have been infeasible for graphs with hundreds of thousands of nodes (20–24). There is a need to quantify and visualise the complexity of genome graphs to shed light on the regions of the genome with salient structural features. Such a method would enable researchers to locate previously unexplored parts of the genome with potential functional significance.

Moreover, past studies that establish methods to work with genome graphs and demonstrate the advantages of genome graphs over the traditional linear reference genome have constructed their reference genome graphs in multiple ways. Each study has augmented a different set of variants onto the linear reference to construct their graph reference without a clear rationale behind the variant set

selection (25). Some studies have augmented the complete set of variants from a prominent species-specific database (26–28). Other studies have filtered out rare variants from the variant set before adding them to the genome graph reference (29–31). The effects of adding rare variants onto the genome graphs are not fully understood yet, and a systematic analysis is required to resolve this issue at the scale of human genomes.

Given the above context on certain opportunities and challenges pertaining to the structural complexity of, and efficient variant calling using genome graphs, the contributions of our study are two-fold:

(i) First, we propose a novel framework to characterise the structural properties of a genome graph. Specifically, our framework helps summarise and visualise the entire human genome graph's structure in a single figure, and identify key genomic loci that exhibit increased individual-to-individual variability. We applied our framework to compare the structures of two human pan-genome graphs, built from 2504 diverse samples in the 1000 Genomes Project: one augmenting only common variants and the other with all variants, including rare ones. We observed that the rare variants increased the variability of the genome graph by 10-fold as expected, and hypervariability by a larger magnitude of 50-fold. Our framework highlighted biologically significant regions of the human genome, like the HLA and DEFB gene loci, as well as several novel regions with increased diversity.

(ii) Second, to call variants using our constructed graphs, we had to optimise the running times of existing genome-graph-based variant calling workflows to the scale of human genome graphs. We then applied our optimised variant calling workflow on a benchmark dataset, and compared the variants called using each of the two constructed graphs – surprisingly, this comparison showed that adding rare variants to genome graphs did not significantly affect the variant calling performance. Additionally, we reiterated that genome graphs were better reference structures than linear references as they captured 9.83% more variants at similar F1 scores.

Taking the above results together, through this work, we demonstrate the capability of our proposed methods to visualise, quantify and compare the intricate genetic complexities within the sample set represented by the genome graphs, while efficiently uncovering novel variants in newly sequenced samples.

METHODS

Construction of human pan-genome graphs

Recent developments using genome graphs as a reference structure for genome analysis have spawned numerous tools for their construction and allow researchers to analyse newly sequenced genomes with genome graphs (32). While most of the existing tools work well for smaller genomes, they rarely perform optimally when the size of the genome scales up. Very few software run smoothly at the scale of human genomes, with *vg toolkit* (14) being one of them. The *vg toolkit* is one of the widely used, openly available tools that can efficiently handle the construction of human genome graphs and analyse whole genome sequences with the same. It is an actively maintained software with an extensive set of functionalities. Based on these characteristics and our aim to use open source software framework to develop tools for genome graphs and their analysis, we used the *vg toolkit (version v1.42.0)* in our study to construct human genome graphs and parts of downstream analyses.

Genome graph construction with the *vg toolkit* involved incorporating variants over a linear reference structure to generate the edges and nodes that comprised the graph-based reference structure. We have used the GRCH38 (hg38) reference genome as the linear reference structure for constructing the genome graph. The 1000 Genomes Project (7), built on the Human Genome Project, sequenced hundreds of individuals from diverse populations and revealed millions of genetic variations. The variants from this project (1KGP variants) captured the genetic heterogeneity of the species well and were used to construct the human pan-genome graph. As the 1KGP variant set called with hg38 reference genome was incomplete for chromosome Y, we lifted over the Y chromosome variants called from hg37 to hg38 and used it to construct and analyse genome graphs.

Human pan-genome graphs can be constructed by augmenting the 1KGP variants onto the hg38 reference. However, most of the variants from the 1KGP project were rare (Table 1), and the added advantage of augmenting such variants onto the genome graph was uncertain. To overcome this lack of resolution, we constructed and compared two types of genome graphs: one with the entire 1KGP variants augmented onto hg38, called the 1KGP Complete Genome Graph, and the other excluding rare variants with alternate allele frequency less than 5%, called the 1KGP Common Genome Graph (Figure 1).

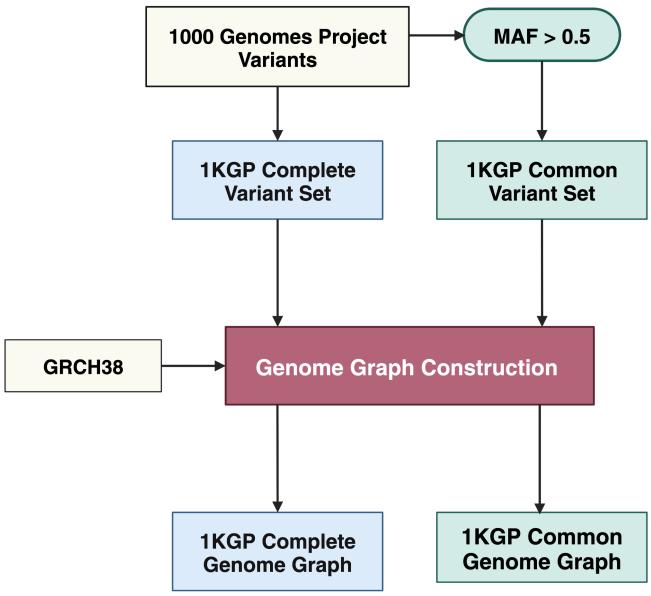

Figure 1: Genome graph construction workflow. '1KGP Complete Genome Graph' augments all the variants from the 1000 genomes project. '1KGP Common Genome Graph' augments only the variants from the 1000 genomes project, with an alternate allele frequency of 5%.

Structural analysis of genome graphs

Genome graphs encased within them a reference path that corresponded to the linear reference genome on which the variants were augmented (14). The human pan-genome graphs constructed in our study encompassed a path that retraced the hg38 reference genome. Nodes in the reference path backtracked on the linear coordinate system, and this property of the genome graph was a bridge between the two reference structures (32). Every variant augmented onto the genome graph created a path that diverged from the reference path at a particular reference node. Each variant path that diverged from a reference node would increase the out-degree of that corresponding reference node by one. This implied that the out-degree of the reference node directly correlated with the complexity of the human genome at that position. This phenomenon was used to credibly quantify the complexity of the entire human genome through the lens of genome graphs. To systematise the structural analysis of genome graphs, we propose the definitions in Box 1.

Reference path: A path in the genome graph corresponding to the linear reference genome used in its construction.
Reference Node: A node present in the reference path.
Variant path: Paths in the genome graph that arise due to the variants augmented during its construction.
Variant node: A node present in a variant path.
Out-Degree: Number of outgoing edges that emanate from a node.
Variable node: A node with a minimum out-degree of two.
Variability: The presence of variable nodes in a region.
Hypervariable node: A node with a minimum out-degree of five. An out-degree of 5 is chosen to distinguish the variable nodes that are more complex than an SNP.
Hypervariability: The presence of hypervariable nodes in a region.
Invariable region: Parts of reference path that are of a size more than 1000 nodes but devoid of any variability.

Box 1: Definitions of terms used in structural analysis of genome graphs.

The genome graphs created in the *vg toolkit* were extracted in the Graphical Fragment Assembly format (GFA) and were incorporated into a custom *Python* program built on top of *NetworkX* (33). *NetworkX* enabled the effortless application of graph algorithms at the scale of human genomes. We subdivided the reference path in the constructed human pan-genome graphs into bins of 10 Mbp. Out-degrees of all the nodes were calculated for the binned subgraphs from both the genome graphs. Variable and hypervariable nodes from each bin were enumerated. A complete panoramic view of the human pan-genome graphs was obtained by collating the count of such nodes throughout the human genome. Circos plots (34) were used to get a bird's-eye view of the genome graphs. The entire workflow consisting of custom scripts in multiple programming languages and other existing open-source tools was automated using the Snakemake workflow management system (35). These views were then used to compare them and identify the key structural similarities and differences in the reference genome graphs. The functional significance of crucial findings from the structural analysis of human pan-genome graphs was studied in-depth.

Variant calling with genome graphs

Genome graphs have been established as a superior method to capture novel variants (17). A major drawback of using genome graphs as a reference structure in variant calling was that they were not easily scalable to the human genome size,

and they required compute resources and runtimes much higher than a linear reference genome. Even though the *vg toolkit* scaled seamlessly to the human genome size, the default variant caller, *vg call (version 1.4.2)* was compute-intensive and time-consuming. As the mapping algorithm gave output in GAM format, employing existing runtime-efficient variant callers that could work only with the well-established BAM format was not straightforward. However, previous works (36) have established a workaround by converting the GAM file format to the BAM file format by projecting the reads mapped to the variant paths of the genome graph to their originating reference nodes. This allowed us to use existing established and runtime-efficient variant calling algorithms with the *vg toolkit*.

In our study, we designed multiple genome-graph-based variants calling workflows by incorporating existing linear genome-based variant callers, namely *GATK Haplotypecaller (version 4.2)* (37), *Bcftools mpileup (version 1.17)* (38), and *FreeBayes (version 1.3.6)* (39). We compared them with the default variant caller *vg call* bundled with the *vg toolkit*. We mapped the raw sequence reads to the 1KGP Common Genome Graph using the runtime-efficient *vg giraffe* algorithm from the *vg toolkit* suite. We converted the output GAM files to BAM format using *vg surject*. The obtained BAM files were processed according to the requirements of the variant caller used. We then compared the performances of these developed workflows by benchmarking them with the acknowledged Genome In a Bottle dataset (GIAB, 34). HG002 was used for the comparison study, and the variants called in each workflow were compared with the high-confidence variant set of HG002 provided by the GIAB consortium. All the developed workflows were run on a machine with 64 cores and 512GB memory to maintain uniformity. The performances of all the developed genome-graph-based variant calling workflows are presented in Table 4. The better pipeline among them was selected based on the runtime of the workflow and the F1 score, which is the harmonic mean of precision and recall from the GIAB benchmarking.

The parallel implementation of the *Freebayes*-based variant calling workflow had a balance of runtime efficiency and good benchmarking scores. Figure 2 describes the optimal genome-graph-based variant-calling pipeline step-by-step. We wrapped this computational pipeline in the Snakemake workflow management system for improved reproducibility and better ease of use in the future (35). We estimated the benefits of adding rare variants to the human genome graph with the finalised genome-graph-based variant calling pipeline. We processed GIAB samples with the 1KGP Common and Complete Genome Graphs. The variant calling capabilities of these human pan-genome graphs were also compared with that of the hg38

reference. A computational pipeline was constructed to call variants with the linear reference genome. *BWA-MEM* was used as the aligner, and the *GATK* suite was used to call and process variants with hg38. The variants called in the seven GIAB samples by these three workflows were compared correspondingly with their high-confidence variant calls provided by the GIAB consortium. *hap.py* (41) was used to benchmark all three pipelines for both SNPs and INDELs.

RESULTS

Overview of human pan-genome graph construction

The prevalence of rare variants in the 1000 Genomes project with an alternate allele frequency of less than 5% was studied for different sample sizes (Table 1). The individuals in each sample size were picked in a stratified random order for population and gender. As the sample size increased, the total number of variants increased steadily, with rare variants constituting the most. With a growing sample size, more variants specific to one or few individuals were called, causing the count of common variants to be stable.

We constructed and compared two genome graphs to understand the benefits of adding the rare variants that were in the majority to the human pan-genome graph. All the 85,123,169 variants identified across 2,504 individuals were augmented onto hg38 to create the 1KGP Complete Genome Graph. After applying a minimum 5% alternate allele frequency filter on these variants, 8,496,706 common variants from these diverse individuals were extracted and augmented onto hg38 to create the 1KGP Common Genome Graph.

Sample Size	**155**	**311**	**626**	**1,252**	**2,504**
1KGP complete variant set	29,244,784	37,419,301	48,490,270	63,876,089	85,123,169
1KGP common Variant set	8,555,382	8,484,132	8,512,275	8,492,781	8,496,706
% of rare variants	70.75	77.33	82.45	86.97	90.02

Table 1: Count of variants in the 1000 genomes project for different sample sizes. The 1KGP complete variant set included all the variants called in that project for the particular sample set. The 1KGP common variant set included only variants with an alternate allele frequency of more than 0.05 and thus were without the rare variants. The sample sets were selected in a stratified random order based on super-population and gender.

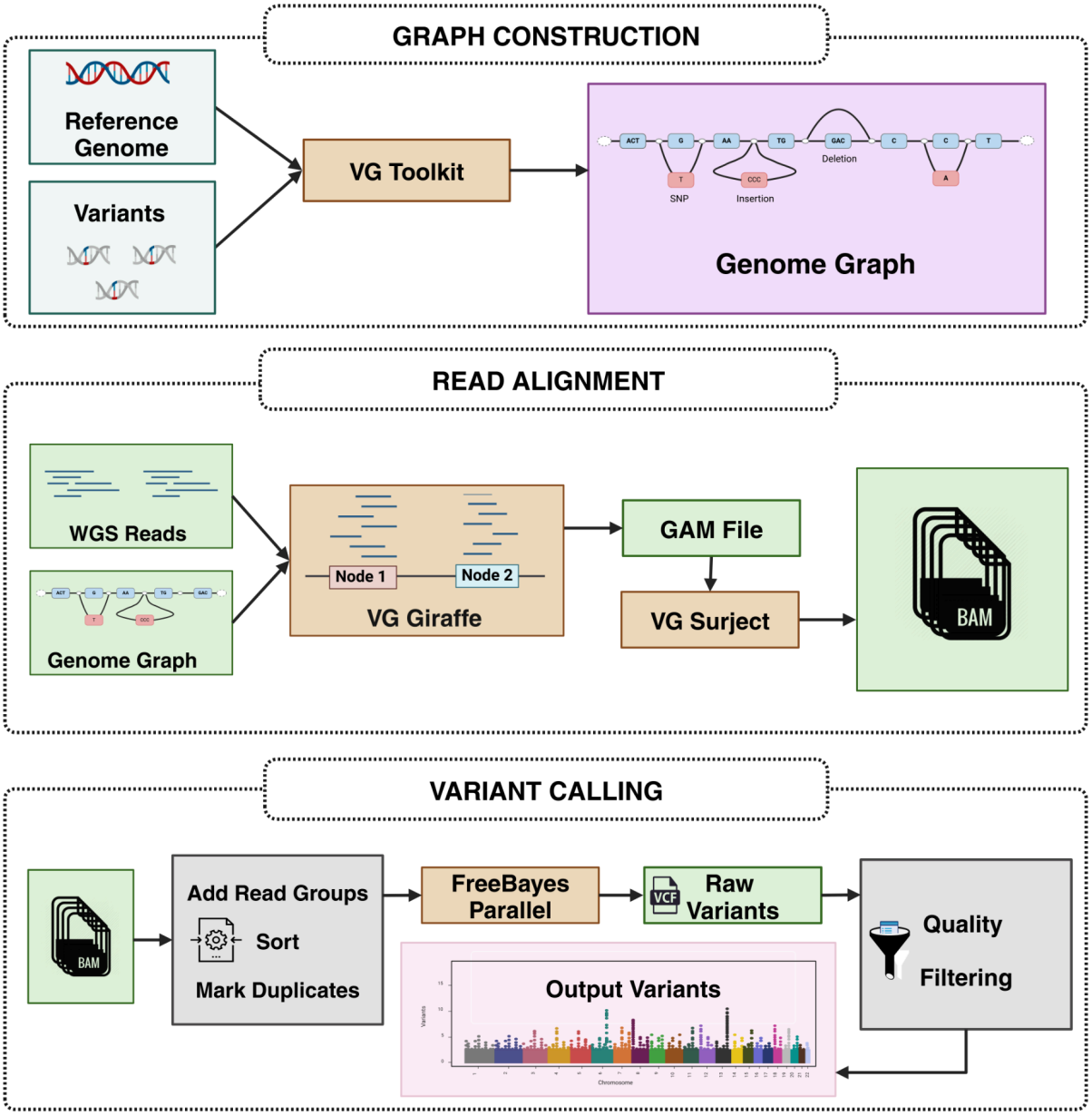

Figure 2: Optimised genome-graph-based variant calling workflow. A step-by-step depiction of the complete genome-graph-based variant calling pipeline that balances performance with runtime. This pipeline was used to compare the advantages of variant-calling performance when adding rare variants onto the genome graphs.

Structural analysis of genome graphs reveals fundamental biological insights

The reference path in the genome graphs was subdivided into bins of window size 10MB. The prevalence of variable and hypervariable regions in each bin in each chromosome was used to get panoramic visualisations of the entire genome graphs. Figure 3 illustrates the bird's-eye view of the human pan-genome graphs.

Interesting structural properties from the human pan-genome graphs can be inferred from Figure 3. The inner track, representing the prevalence of variable nodes, was capped at 524K for the 1KGP Complete Genome Graph, whereas it was 61K for the 1KGP Common Genome Graph. This result was expected as the rare variants were removed in constructing the 1KGP Common Genome Graph. Hence, the variability of the common graph would be lesser when compared to the 1KGP Complete Genome Graph. However, the contours of the bars capturing the distribution of variable nodes in the inner track were similar for both the genome graphs. The Pearson correlation coefficient for the count of variable nodes in the corresponding bins of both the genome graphs was 0.94, indicating a robust linear relationship in their genomic variability. This further implies that the polymorphism in the human genome that arose due to common variants alone retained certain genome topography from the polymorphism that emerged from the entire set of genetic variants.

However, the contours of the bars in the outer track capturing the occurrence of hypervariable nodes were starkly different for both the genome graphs. A Pearson correlation coefficient of 0.11 and a Spearman correlation coefficient of 0.21 were observed for the count of hypervariable nodes in the corresponding bins of both the genome graphs, indicating the absence of a straightforward relationship in their hypervariability. In the 1KGP Complete Genome Graph, hypervariable nodes were present throughout, with the majority of them located in chromosome X. Whereas, the 1KGP common Genome Graph was depleted of hypervariable nodes all over the genome, even in chromosome X. Rare variants that were removed from the 1KGP Common Genome Graph, were mostly driving the hypervariability in the genome. This could mean that the hypervariability observed in the human genome, particularly in chromosome X, originated from different populations independently, as the common genetic variants present across populations did not cause the structural complexity to the same level.

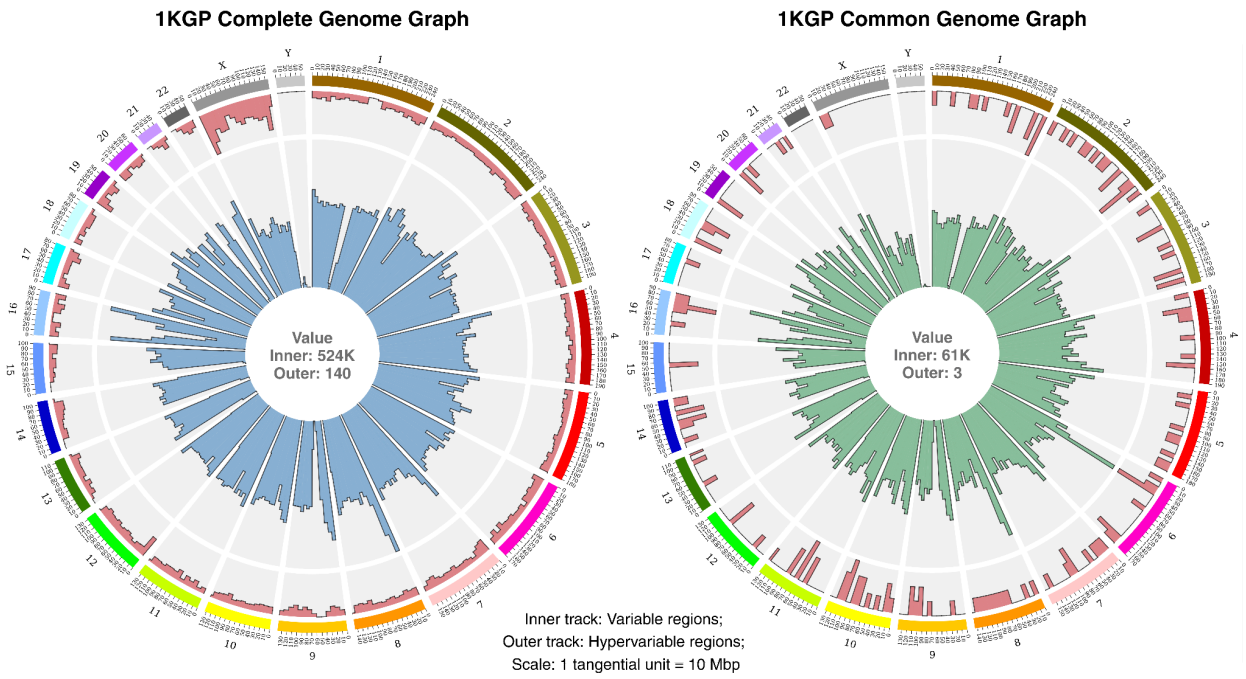

Figure 3: Bird's-eye view of human pan-genome graphs. Each sector in the Circos represents a chromosome. Each tangential bar spans a 10MB genomic window. The number of hypervariable nodes in each bin is depicted in the outer track. The number of variable nodes in each bin is represented in the inner tracks. The text at the centre indicates the maximum value for the inner and outer tracks.

The location of the bins with a high frequency of variable nodes in the two genome graphs is shown in Table 2. A 10MB bin in chromosome 6 from 30-40 Mbp showed the highest variability only in the 1KGP Common genome graph. This bin encompassed most of the Major Histocompatibility Complex (MHC) region, also known as the Human Leukocyte Antigen (HLA) region, of the human genome. Previous studies have established that the HLA region contains one of the most polymorphic gene clusters of the entire human genome. The presence of multiple HLA alleles in the population ensures that at least some individuals within a population will be able to recognise protein antigens produced by any microbe, thus reducing the likelihood that a single pathogen can evade host defences in all individuals in a given species (42). Table 2 also shows that the bin containing HLA had a lower frequency of variable nodes in the 1KGP Complete genome graph when compared to its other bins. However, the same bin included the highest number of variable nodes in the IKGP Common genome graph. This implied that a notable portion of the variants driving polymorphism in HLA were not rare and were shared by individuals in different human subpopulations.

In the human pan-genome graphs, other genomic regions with high variability were observed in chromosomes 8 and 16. The first 10MB bin from chr8 with the highest variability in 1KGP Complete graph encompassed parts of the β-defensin gene clusters called DEFB. DEFB was previously known to be highly variable and was associated with evolutionary importance and disease risk (43). Chromosome 16 enclosed a couple of highly variable bins in both human pan-genome graphs as it hosted several large polymorphisms, often associated with segmental duplications (44). As these bins were highly variable in both the genome graphs, it implied that the polymorphism in these regions originates from common and rare variants. While some of the genomic regions highlighted by this diversity analysis coincided with previously known zones of biological significance, the remaining regions in Table 2 and Figure 3 can be viewed as novel candidate loci, whose diversity could be analysed further to ascertain their importance.

Structural analysis enables the quantification of the complexity of genome graphs

Genome graphs capture the complexity and diversity of a species, and analysing its structure can be crucial for understanding specific aspects of its genome. However, the structure of genome graphs is dynamic and depends on the set of variants augmented during its construction. The number of variable and hypervariable nodes effectively quantifies the genomic complexity encompassed by the genome graph. Nodes with an out-degree of at least two originate at the sites of variants in the genome. So, the number of variable nodes in a genome graph characterises the number of variants augmented at the time of its construction.

Genome Graph	Rank	Region	Variability	Remark
Complete	1	chr8:0-10MB	523,905	Contains parts of the β-defensin gene clusters
Complete	2	chr16: 80-90MB	492,481	Prone to structural variations and duplications
Complete	3	chr16:0-10MB	477,967	Prone to structural variations and duplications
Complete	4	chr8:10-20MB	457,413	Associated with an immune response (BLK)
Complete	5	chr7:0:10MB	429,860	-
Complete	6	chr9:0:10MB	426,628	-
Complete	7	chr4:0:10MB	412,151	Associated with neurological development (HTT)

Complete	8	chr16:70:80MB	400,705	Prone to structural variations and duplications
Complete	9	chr9:10-20MB	396,851	-
Complete	10	chr3:0-10MB	387,369	-
Common	1	chr6:30-40MB	60,793	Overlaps with MHC (Major Histocompatibility Complex) region
Common	2	chr8:0-10MB	54,037	Contains parts of the β-defensin gene clusters
Common	3	chr16:80-90MB	49,302	Prone to structural variations and duplications
Common	4	chr8:10-20MB	48,197	Associated with an immune response (BLK)
Common	5	chr16:0-10MB	44,329	Prone to structural variations and duplications
Common	6	chr4:180-190MB	43,360	-
Common	7	chr7:0:10MB	42,340	-
Common	8	chr3:0-10MB	41,827	-
Common	9	chr9:0:10MB	41,775	-
Common	10	chr4:0:10MB	41,037	Associated with neurological development (HTT)

Table 2: Top ten bins with the most variability in each of the human pan-genome graphs. The 1KGP Complete Genome Graph is abbreviated by 'Complete' and the 1KGP Common Genome Graph is abbreviated as 'Common'. Variability is calculated as the count of variable nodes in the corresponding bin.

The maximum possible out-degree of a node representing an SNP in the genome graphs was four (one reference and three possible alternate alleles). Intending to capture the highly complex parts of the genome, we defined hypervariable nodes as more complex than SNP nodes. Hypervariable nodes, representing genomic regions with high levels of polymorphism, can be essential for studying population-specific genetic diversity and capturing regions of the genome that substantially impact various biological processes and functional relationships. These hypervariable nodes can potentially be used as markers that can distinguish different genome graphs built for the same species, as other graph-derived metrics like the number of nodes and edges cannot achieve this.

Table 3 compares the structural complexity of human genome graphs using the proposed framework. Variability and hypervariability of the genome graph were quantified by the total counts of variable and hypervariable nodes, respectively. The effect of rare variants on the structural complexity is computed as the fold change observed in the 1KGP Complete Genome Graph in the corresponding complexity types. It can be noted that the variability of the human pan-genomes represented the count of variants used in its construction, and it closely followed the variant set sizes observed for 2,504 samples (Table 1). The count of variable nodes was higher than the count of variants due to the possibility that alleles of INDELs with similar sequences were simplified, giving rise to more variable nodes. The 10-fold change in variability caused by the rare variants also complied with the fact that nearly 90% of the variants observed in the 2,504 samples were rare. It was also observed that rare variants have more effect on hypervariability, resulting in a 50-fold change between the two genome graphs.

Complexity Type	**1KGP Complete Genome Graph**	**1KGP Common Genome Graph**	**Fold Change by rare variants**
Variability	89,352,596	8,551,031	10.45
Hypervariability	7,446	148	50.31

Table 3: Measure of variability and hypervariability in both the human pan-genome graphs. Fold change was computed for the counts in the 1KGP Common Genome Graph.

The degree-based distribution of hypervariable nodes in the constructed human pan-genomes is shown in Figure 4. As expected, the 1KGP Common Genome Graph had restricted distribution of hypervariability not only for genomic position but also with degree. All the 148 hypervariable nodes observed in the 1KGP Common Genome Graph had an out-degree of either five or six. On the other hand, the hypervariability was well distributed in terms of both degree and genomic position in the 1KGP Complete Genome Graph. Out of the 7,446 hypervariable nodes observed, two nodes were seen with the maximum out-degree spanning twelve. When the exact location of the most hypervariable nodes was determined, it was found that both these nodes were adjacent to each other and were present in chromosome 1. Figure S1 depicts the path-level representation of the adjoining 12-degree nodes in the 1KGP Complete genome graph obtained using the default visualisation command bundled with the *vg toolkit*.

The human genome is comprised of large zones of invariance

Invariant regions are essential components of the genome, reflecting the conservation of genomic segments across different individuals, populations, or species. While invariant zones across individuals suggest a critical role in maintaining essential biological functions, invariance across the genomes of species helps infer evolutionary relationships and divergence patterns among different organisms. Identifying and studying invariant zones in the genome can contribute to various aspects of genomics, including evolutionary biology, functional genomics, and biomedical research.

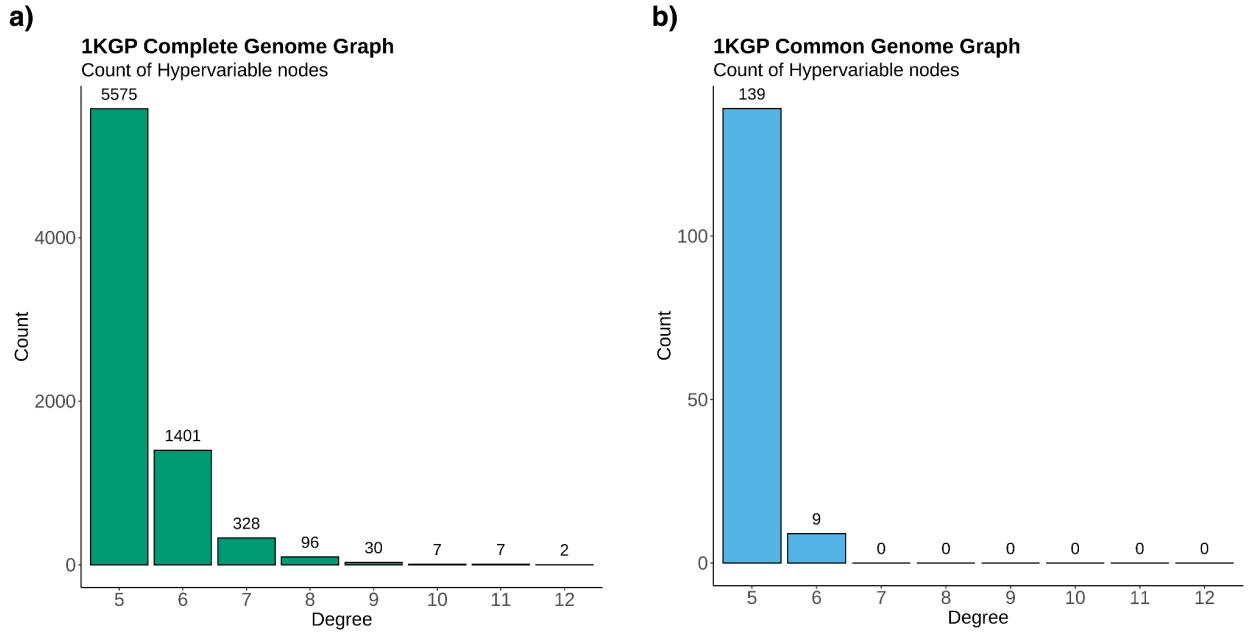

Figure 4. Hypervariable nodes in human pan-genome graphs

(a) Degree distribution of the hypervariable nodes in the 1KGP Complete Genome Graph.

(b) Degree distribution of the hypervariable nodes in the 1KGP Common Genome Graph.

Genome graphs, apart from emphasising hypervariable nodes, also offer a convenient way to discover the invariant zones in the genome. In our study, we defined invariant zones as the regions in the genome graph composed of at least 1000 nodes devoid of any variant paths. This definition translated to a stretch of the genome at least 32 Kbp in size, with no variants found in the samples included in the study. As the 1000 genome project was a very diverse cohort, invariable regions

identified in the human pan-genome graphs were conserved for individuals across diverse populations.

The 1KGP Common Genome Graph was created by removing the rare variants from the variable regions of the 1KGP Complete Genome Graph. Removal of the rare variants reduced the number of variable nodes from the genome graph and, depending on the genomic location, removal could have three possible effects: i) when variable nodes in a region were removed enough, new invariant zones were created; ii) when variable nodes present at the edges of existing invariant zones were cleared, the length of the corresponding invariant zone increased; iii) when variable nodes were not reduced enough in a region, no change was observed for the invariant zones.

The distribution of invariant regions in both the human pan-genome graphs is presented in Figure 5. It can be inferred from Figure 5a, b that all three effects had an effect when rare variants were removed from the genome graph.

The number of invariant zones was unchanged between the two genome graphs for twelve chromosomes, but the aggregate invariability differed for all the chromosomes. This implied that only the lengths of existing invariant regions have increased in these chromosomes due to the removal of rare variants. In these chromosomes, all the variable node removal at the edges of existing invariant zones increased their lengths, and removing other variable nodes was not enough to create new invariant zones. Meanwhile, the reduction of the median length of invariant zones in the 1KGP Common Genome Graph, for example, in chromosomes 2, 7, 16 and 19, implied the creation of newer invariant zones that were of smaller lengths than the ones observed in the 1KGP Complete Genome Graph. Sizable invariant zones with more than 1 MBbp lengths were observed at various regions of both genome graphs (Figure S2).

Optimisation of existing genome-graph-based variant calling workflows

To overcome the extensive computational requirements of the genome-graph-based variant calling pipelines, we developed multiple workflows that employed linear-genome-based variant callers such as *Bcftools mpileup*, *GATK Haplotypecaller* and *Freebayes-parallel*, in conjunction with genome-graph-mapped files. Therefore, we mapped the GIAB sample HG002 to the 1KGP Common Genome Graph using the *vg giraffe* algorithm and converted the obtained GAM file to the BAM format. We also called variants using the variant caller *vg call* bundled with the *vg toolkit* and compared the results with the other developed workflows.

The variants generated from each workflow were compared with the high-confidence variant calls published by the GIAB consortium. *hap.py* was used to benchmark these developed genome-graph-based workflows.

The metrics quantifying the performance of developed variant-calling workflows are summarised in Table 4. The runtimes of each post-graph-alignment step for the developed workflows are presented in detail in Table S1a-d. It can be seen that using *Bcftools mpileup* and *Freebayes* significantly reduced the computing time of the genome-graph-based variant calling. The sharp increase in the runtime of the *GATK HaplotypeCaller* pipeline was due to the extensive preprocessing steps. As reflected in the F1 scores, particularly in the case of INDELs, *GATK HaplotypeCaller* was the best-performing workflow. However, the parallel implementation of *Freebayes* scores was very similar to F1 scores and was the fastest pipeline. Considering the balance of performance and runtime, we used the *Freebayes-parallel* workflow for our downstream analyses.

Variant Caller	Count of SNPs	F1 score SNPs	Count of INDELs	F1 score INDELs	VC Runtime (HH: MM)
vg call	**4,538,504**	0.924	**1,207,055**	0.481	15:24
GATK HaplotypeCaller	3,922,180	**0.972**	956,122	**0.866**	36:46
Bcftools mpileup	3,979,841	0.971	1,036,455	0.474	06:50
Freebayes-parallel	3,980,611	**0.972**	851,680	0.865	**04:51**

Table 4: Metrics derived from the variants called by different variant callers. VC runtime is the compute time each workflow takes after the alignment step. The best values are highlighted in bold.

Genome graphs are better reference structures than linear genomes for variant identification

GIAB samples were processed with both the 1KGP Common Genome Graph and the 1KGP Complete Genome Graph to understand better the variant calling benefits of adding rare variants to the human pan-genome. The *Freebayes*-based variant calling workflow that benchmarked well with HG002 was used to process all the

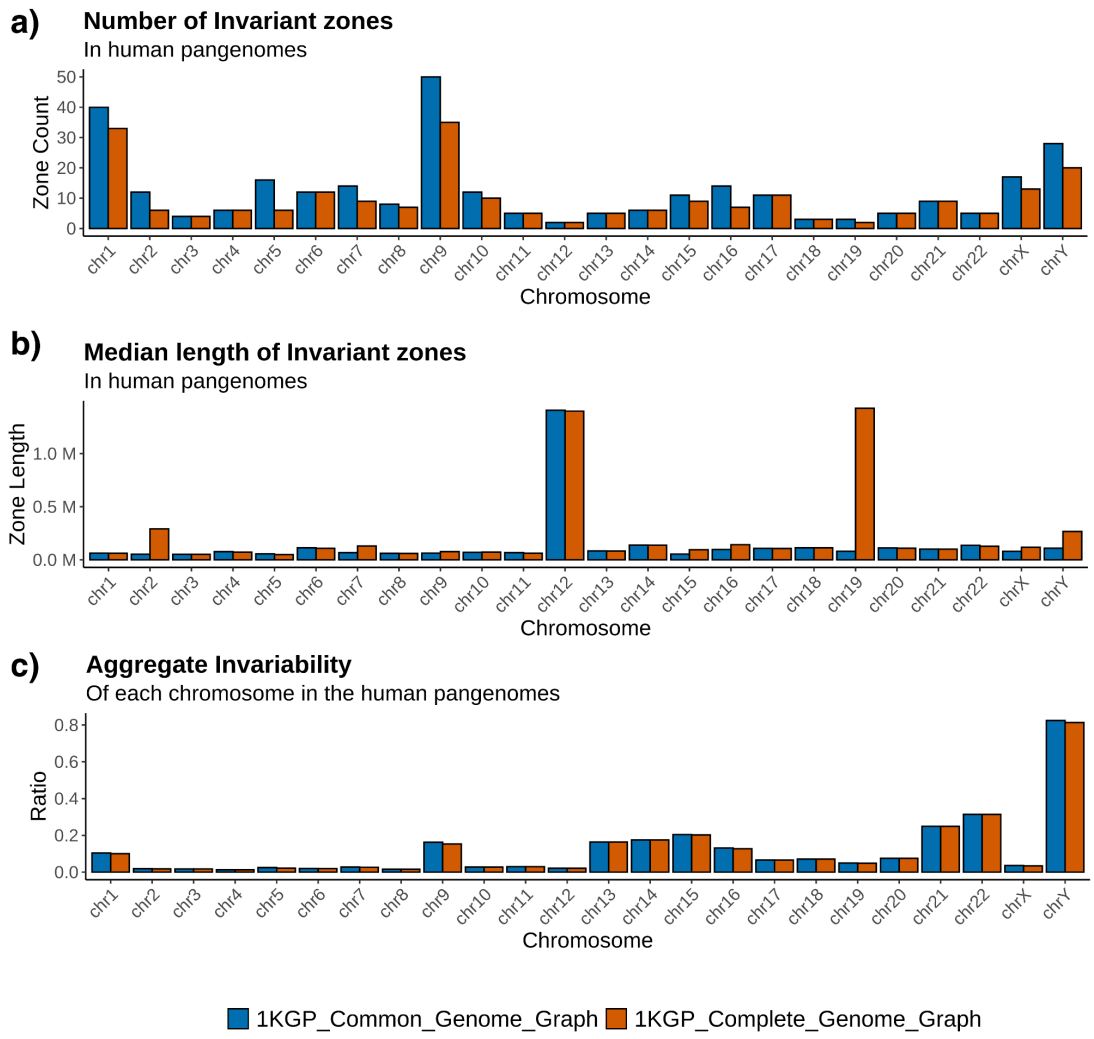

Figure 5: The distribution of invariant regions in human pan-genome graphs

(a) The total number of invariant zones in each chromosome for the two human pan-genomes constructed. The y-axis represents the count of invariable zones and the x-axis represents chromosomes. For both the genome graphs, chromosomes 3, 4, 6, 11, 12, 13, 14, 17, 18, 20, 21, 22 have the same number of invariant regions.

(b) The median length of the invariant zones in each chromosome. The y-axis represents the length of the zone in base pairs and the x-axis represents the chromosomes. The length of the zones in BP is calculated as (number of nodes in the zone) times 32.

(c) The volume of chromosomes that is invariant in the human pan-genome graphs. The y-axis represents the invariability, and the x-axis represents the chromosomes. Invariability is calculated as (Total length of invariable zones in a chromosome in BP)/(Total length of the chromosome in BP).

GIAB samples with both the constructed genome graphs. We also processed the samples with the linear hg38 with the established *BWA-GATK* workflow to compare different reference structures. The metrics extracted from the variants called in all three workflows are given in Figure 6.

The genome graph referenced workflows captured, on average, 434K more variants than the linear reference workflow. This meant the optimised genome-graph-based

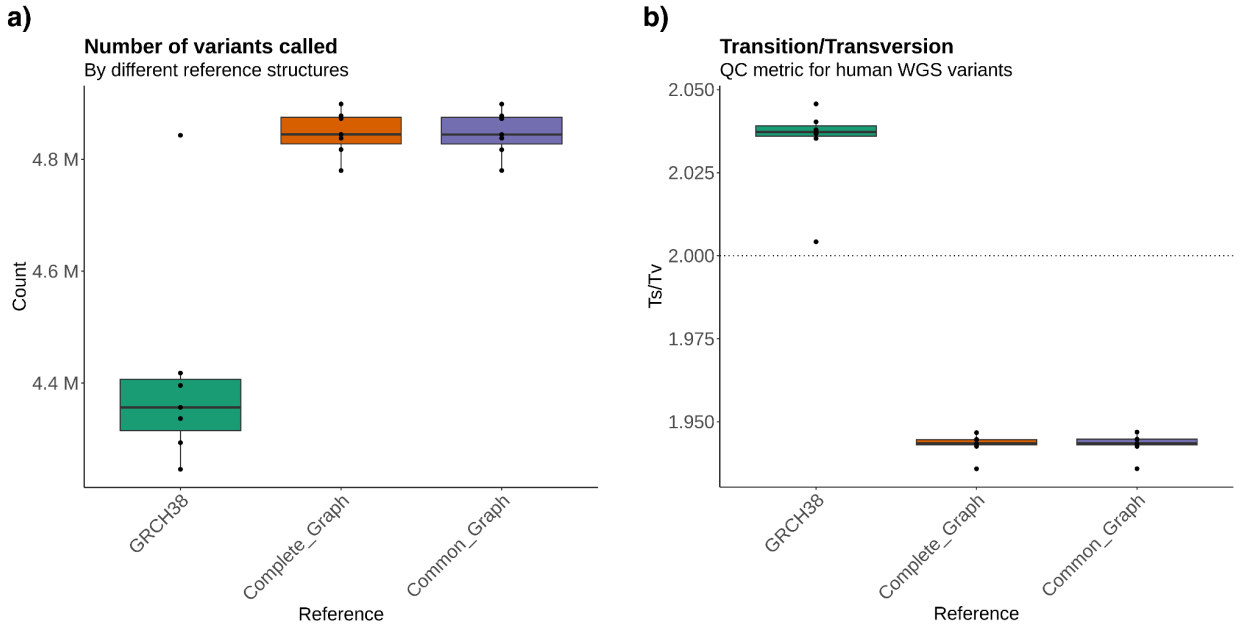

Figure 6: Variant calling metrics for different reference structures: The 1KGP Complete Genome Graph is represented as Complete_Graph and the 1KGP Common Genome Graph is represented as Common_Graph in the x-axis of both the figures.

(a) The total number of variants captured in all the seven GIAB samples by the three reference structures. *BWA-GATK* was used to process samples with GRCH38, and the finalised *vg-giraffe-Freebayes-parallel* was used to process samples with the genome graphs.

(b) Distribution for the ratio of the number of transitions to the number of transversions for the variants called in GIAB samples by the different reference structures.

variant calling workflows captured 9.83% more variants than the linear-reference-based workflow. Examining the variants closer, we observed that in the case of SNPs, the genome graph workflows consistently outperformed the linear workflow, whereas in the case of INDELs, in HG002 alone, the linear pipeline called more of that variant type (Figure S3). To determine statistical significance, a one-tailed paired sample Wilcoxson signed rank test was performed. The 1KGP

Complete Genome Graph (p-value = 0.0078) and the 1KGP Common Genome Graph (p-value = 0.0078) reference structures captured more variants than the linear hg38. Regarding the quality control metric, all three reference structures achieved a Ts/Tv ratio of nearly 2.0, an expected threshold for the human whole genome variants (45).

However, adding rare variants to the genome graph did not give any significant improvements in terms of the number of variants called. The 1KGP Complete Genome Graph that augmented the rare variants from diverse individuals could not capture more variants (p-value = 0.078) than the 1KGP Common Genome Graph. The average increase in the variant count obtained by the more complex genome graph was only 105. In HG005 and HG006, the less complex 1KGP Common Genome Graph was observed to capture slightly more variants than the 1KGP Complete Genome Graph, which augmented all variants.

To verify if the new variants captured by the genome graphs were not false positives, we benchmarked the variants called by all the reference structures with the high-confidence variants set. Except for HG002, the F1 scores for the SNPs and INDELs were higher for the genome-graph-based workflows (Figure S4, Table S2). We observed that for both SNPs and INDELs, the 1KGP Common Genome Graph (p-value = 0.0156 for both SNPs and INDELs) and the 1KGP Complete Genome Graph (p-value = 0.0156 for both SNPs and INDELs) had higher F1 scores than for the linear genome-referenced workflow (Figure S4, Table S2).

Even though it was unclear from the visualisation, the F1 score for SNPs was observed to be higher for the 1KGP Common Genome Graph by an average of 1.2 x e^{-05} than the 1KGP Complete Genome Graph (p-value = 0.0078). The difference between the genome graphs was not significant (p-value = 0.0711) for the F1 scores of INDELs. Considering the number of variants and the benchmarking results, the linear reference genome-based workflow underperformed the genome-graph-based workflows. As the performance of the 1KGP Common Genome Graph was at par with that of the 1KGP Complete Genome Graph, no benefits in terms of variant calling were observed while adding rare variants to the human genome graph.

DISCUSSION

In this study, we designed a novel framework to represent, capture and thoroughly analyse the structural properties of genome graphs and uncover the underlying genomic complexity of the populations or species they represent. The proposed structural analysis method effectively scales to the whole human genome size and

allows one to visualise the entire human genome graph simultaneously. Such a view can be a foundation for visualising, navigating, and prioritising sites of interest from the human genome, which is valuable for in-depth research. We segregated the complex regions of the genome graphs into three broad categories: variable, hyper-variable and invariable. While the study of variable and hypervariable regions is essential for understanding the genetic makeup of populations and exploring the functional implications of genetic diversity, invariant zones of the genome help identify conserved regions with high evolutionary significance.

We applied our framework to analyse, visualise and compare the structural properties of two human pan-genome graphs constructed in this study. When rare variants are augmented to the genome graph, the aggregate invariability of the genome graph decreases. We also observed that the addition of rare variants increased the variability of the genome graph by 10-fold and the hypervariability by 50-fold. The framework highlighted biologically significant regions of the genome, such as HLA and DEFB (Table 2), demonstrating the potential of our proposed method to unearth meaningful genomic loci. When a similar analysis is performed for genomes of novel populations, human or otherwise, we could potentially discover critical regions of the genome affecting the organism's diversity, health and evolution. The structural analysis framework of genome graphs can also be advantageous in capturing the genomic complexity of a novel cohort, effectively paving a new way to compare genetic diversity quantitatively. Population-level studies could benefit from such a comparison, as it highlights the similarities and differences in the polymorphism patterns between populations. Polymorphic regions in the genome, which vary between populations, can influence disease susceptibility and drug response. Identifying these regions is key to advancing personalised medicine (46). Thus, the proposed framework for structural analysis of genome graphs is vital for both fundamental biological research, and the development of strategies for preventing and treating genetic diseases.

As the existing genome-graph-based variant calling workflow was running time intensive, we tested different existing variant callers in conjunction with the *vg toolkit* and identified the optimal pipeline. Using this pipeline, we confirmed the superior variant-calling performance of genome graphs using the GIAB dataset, as genome graphs captured 434K more variants than linear-genome workflows while achieving higher F1 scores. When applied to newly sequenced human genomes, this optimised workflow can identify novel variants that the existing linear variant callers would miss out on. However, incorporating rare variants into the human genome graph showed no significant advantage in variant calling, as the 1KGP

Common Genome Graph performed similarly to the 1KGP Complete Genome Graph, capturing only 105 more variants, with the former achieving better SNP F1 scores. When compared to graphs with simpler structures, the underperformance of high-complexity genome graphs aligns with previous studies (25) that evaluated the read-mapping performance of genome graphs with varying complexities. Through our study, we extended this comparison to the variant calling performance of genome graphs and validated it on the scale of the human whole genome.

When analysing the variants captured in the 1000 genomes project for different stratified sample sets, we identified that with increasing sample size, the percentage of rare variants increased while the number of common variants remained almost constant. So, constructing an extensive human genome graph with only the common variants from diverse individuals can be achieved with a smaller sample size. In contrast, constructing a well-representative genome graph that augments the rare variants would require a large sample size. As population-level studies are done in phases (17), and fewer samples are sequenced at the beginning of the study, analysis of the new genomes can benefit from the construction of a population-specific genome graph that augments only common variants that can stay stable with increasing sample size when compared to all the variants that expand constantly. Nevertheless, a detailed variant count analysis specific to the population is required before such a population-specific genome-graph-based analysis begins.

The framework for the structural analysis of genome graphs elucidated in this paper is coded for and tested on the genome graphs constructed with the *vg toolkit*. However, the proposed idea can be extended to variational genome graphs built with other software, as the structural analysis framework is pivoted only around the traversal of the reference path. In this study, we have used hg38 instead of the latest T2T (47) as the reference genome on which the variant paths were added to create the genome graphs. The formulated structural analysis framework would function well, irrespective of the reference genome. Moreover, this proposed idea is not just limited to human genome graphs but can extend smoothly to genome graphs representing the genetic complexity in other species.

In terms of future directions, the annotation of genome graphs with meta-information can complement the proposed structural analysis well and enable the application of more sophisticated graph algorithms to genome graphs. Employing these methods to population-specific human genome graphs allows for a comprehensive assessment of genetic diversity, population structure, disease associations, and evolutionary dynamics within and across populations. This

information can enhance our understanding of the diversity in human genetics and how it impacts health, diseases, and the historical dynamics of populations.

SOFTWARE AND CODE AVAILABILITY

Both the optimised genome-graph-based variant calling pipeline and the novel structural analysis workflow are automated with Snakemake for better usability and reproducibility. These codes are available at https://github.com/IBSE-IITM/HumanGenomeGraphs. The *BWA-GATK* workflow is available at https://github.com/IBSE-IITM/genome-analysis-pipeline.

ACKNOWLEDGEMENTS

We thank Philge Philip, Sai Sruthi Amirtha Ganesh, Keerthika Moorthy, Harshita Agarwal, and Centre for Integrative Biology and Systems Medicine (IBSE) and Robert Bosch Centre for Data Science and Artificial Intelligence (RBCDSAI) members for their valuable discussions and comments.

FUNDING INFORMATION

The work was supported by the Department of Biotechnology, Govt. of India (BT/GenomeIndia/2018) and the Centre for Integrative Biology and Systems Medicine, IIT Madras (BIO/18-19/304/ALUM/KARH). MN was supported by the Wellcome Trust/DBT grant IA/I/17/2/503323.

Conflict of interest statement. None declared.

REFERENCES

1. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W., *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
2. International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.
3. Homo sapiens genome assembly GRCh38 *NCBI*.
4. Schneider,V.A., Graves-Lindsay,T., Howe,K., Bouk,N., Chen,H.-C., Kitts,P.A., Murphy,T.D., Pruitt,K.D., Thibaud-Nissen,F., Albracht,D., *et al.* (2017) Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.*, **27**, 849–864.
5. Green,R.E., Krause,J., Briggs,A.W., Maricic,T., Stenzel,U., Kircher,M., Patterson,N., Li,H., Zhai,W., Fritz,M.H.-Y., *et al.* (2010) A Draft Sequence of the Neandertal Genome. *Science*, **328**, 710–722.
6. Reich,D., Nalls,M.A., Kao,W.H.L., Akylbekova,E.L., Tandon,A., Patterson,N., Mullikin,J., Hsueh,W.-C., Cheng,C.-Y., Coresh,J., *et al.* (2009) Reduced Neutrophil Count in

People of African Descent Is Due To a Regulatory Variant in the Duffy Antigen Receptor for Chemokines Gene. *PLOS Genet.*, **5**, e1000360.
7. Auton,A., Abecasis,G.R., Altshuler,D.M., Durbin,R.M., Abecasis,G.R., Bentley,D.R., Chakravarti,A., Clark,A.G., Donnelly,P., Eichler,E.E., *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
8. Lek,M., Karczewski,K.J., Minikel,E.V., Samocha,K.E., Banks,E., Fennell,T., O'Donnell-Luria,A.H., Ware,J.S., Hill,A.J., Cummings,B.B., *et al.* (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, **536**, 285–291.
9. Wall,J.D., Stawiski,E.W., Ratan,A., Kim,H.L., Kim,C., Gupta,R., Suryamohan,K., Gusareva,E.S., Purbojati,R.W., Bhangale,T., *et al.* (2019) The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature*, **576**, 106–111.
10. Karczewski,K.J., Francioli,L.C., Tiao,G., Cummings,B.B., Alföldi,J., Wang,Q., Collins,R.L., Laricchia,K.M., Ganna,A., Birnbaum,D.P., *et al.* (2020) The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, **581**, 434–443.
11. Ballouz,S., Dobin,A. and Gillis,J.A. (2019) Is it time to change the reference genome? *Genome Biol.*, **20**, 1–9.
12. Paten,B., Novak,A.M., Eizenga,J.M. and Garrison,E. (2017) Genome graphs and the evolution of genome inference. *Genome Res.*, **27**, 665–676.
13. Ameur,A. (2019) Goodbye reference, hello genome graphs. *Nat. Biotechnol.*, **37**, 866–868.
14. Garrison,E., Sirén,J., Novak,A.M., Hickey,G., Eizenga,J.M., Dawson,E.T., Jones,W., Garg,S., Markello,C., Lin,M.F., *et al.* (2018) Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.*, **36**, 875–879.
15. Sirén,J. (2017) Indexing Variation Graphs. In *2017 Proceedings of the Meeting on Algorithm Engineering and Experiments (ALENEX)*, Proceedings. Society for Industrial and Applied Mathematics, pp. 13–27.
16. Liao,W.-W., Asri,M., Ebler,J., Doerr,D., Haukness,M., Hickey,G., Lu,S., Lucas,J.K., Monlong,J., Abel,H.J., *et al.* (2023) A draft human pangenome reference. *Nature*, **617**, 312–324.
17. Tetikol,H.S., Turgut,D., Narci,K., Budak,G., Kalay,O., Arslan,E., Demirkaya-Budak,S., Dolgoborodov,A., Kabakci-Zorlu,D., Semenyuk,V., *et al.* (2022) Pan-African genome demonstrates how population-specific genome graphs improve high-throughput sequencing data analysis. *Nat. Commun.*, **13**, 4384.
18. Rakocevic,G., Semenyuk,V., Lee,W.-P., Spencer,J., Browning,J., Johnson,I.J., Arsenijevic,V., Nadj,J., Ghose,K., Suciu,M.C., *et al.* (2019) Fast and accurate genomic analyses using genome graphs. *Nat. Genet.*, **51**, 354–362.
19. Andreace,F., Lechat,P., Dufresne,Y. and Chikhi,R. (2023) Comparing methods for constructing and representing human pangenome graphs. *Genome Biol.*, **24**, 274.
20. Wick,R.R., Schultz,M.B., Zobel,J. and Holt,K.E. (2015) Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics*, **31**, 3350–3352.
21. Beyer,W., Novak,A.M., Hickey,G., Chan,J., Tan,V., Paten,B. and Zerbino,D.R. (2019) Sequence tube maps: making graph genomes intuitive to commuters. *Bioinformatics*, **35**, 5318–5320.
22. Kunyavskaya,O. and Prjibelski,A.D. (2019) SGTK: a toolkit for visualization and assessment of scaffold graphs. *Bioinformatics*, **35**, 2303–2305.
23. Gonnella,G., Niehus,N. and Kurtz,S. (2019) GfaViz: flexible and interactive visualization of GFA sequence graphs. *Bioinformatics*, **35**, 2853–2855.
24. Guarracino,A., Heumos,S., Nahnsen,S., Prins,P. and Garrison,E. (2022) ODGI:

understanding pangenome graphs. *Bioinformatics*, **38**, 3319–3326.
25. Pritt,J., Chen,N.-C. and Langmead,B. (2018) FORGe: prioritizing variants for graph genomes. *Genome Biol.*, **19**, 1–16.
26. Danek,A., Deorowicz,S. and Grabowski,S. (2014) Indexes of Large Genome Collections on a PC. *PLOS ONE*, **9**, e109384.
27. Liu,B., Guo,H., Brudno,M. and Wang,Y. (2016) deBGA: read alignment with de Bruijn graph-based seed and extension. *Bioinformatics*, **32**, 3224–3232.
28. Schneeberger,K., Hagmann,J., Ossowski,S., Warthmann,N., Gesing,S., Kohlbacher,O. and Weigel,D. (2009) Simultaneous alignment of short reads against multiple genomes. *Genome Biol.*, **10**, 1–12.
29. Eggertsson,H.P., Jonsson,H., Kristmundsdottir,S., Hjartarson,E., Kehr,B., Masson,G., Zink,F., Hjorleifsson,K.E., Jonasdottir,A., Jonasdottir,A., *et al.* (2017) Graphtyper enables population-scale genotyping using pangenome graphs. *Nat. Genet.*, **49**, 1654–1660.
30. Kim,D., Paggi,J.M., Park,C., Bennett,C. and Salzberg,S.L. (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.*, **37**, 907–915.
31. Maciuca,S., del Ojo Elias,C., McVean,G. and Iqbal,Z. (2016) A Natural Encoding of Genetic Variation in a Burrows-Wheeler Transform to Enable Mapping and Genome Inference. In Frith,M., Storm Pedersen,C.N. (eds), *Algorithms in Bioinformatics*, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 222–233.
32. Eizenga,J.M., Novak,A.M., Sibbesen,J.A., Heumos,S., Ghaffaari,A., Hickey,G., Chang,X., Seaman,J.D., Rounthwaite,R., Ebler,J., *et al.* (2020) Pangenome Graphs. *Annu. Rev. Genomics Hum. Genet.*, **21**, 139–162.
33. Hagberg,A., Swart,P.J. and Schult,D.A. (2008) Exploring network structure, dynamics, and function using NetworkX Los Alamos National Laboratory (LANL), Los Alamos, NM (United States).
34. Krzywinski,M., Schein,J., Birol,I., Connors,J., Gascoyne,R., Horsman,D., Jones,S.J. and Marra,M.A. (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.*, **19**, 1639–1645.
35. Mölder,F., Jablonski,K.P., Letcher,B., Hall,M.B., Tomkins-Tinch,C.H., Sochat,V., Forster,J., Lee,S., Twardziok,S.O., Kanitz,A., *et al.* (2021) Sustainable data analysis with Snakemake. 10.12688/f1000research.29032.1.
36. Sirén,J., Monlong,J., Chang,X., Novak,A.M., Eizenga,J.M., Markello,C., Sibbesen,J.A., Hickey,G., Chang,P.-C., Carroll,A., *et al.* (2021) Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science*, **374**, abg8871.
37. McKenna,A., Hanna,M., Banks,E., Sivachenko,A., Cibulskis,K., Kernytsky,A., Garimella,K., Altshuler,D., Gabriel,S., Daly,M., *et al.* (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.
38. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R., and 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
39. Garrison,E. and Marth,G. (2012) Haplotype-based variant detection from short-read sequencing. 10.48550/arXiv.1207.3907.
40. Krusche,P., Trigg,L., Boutros,P.C., Mason,C.E., De La Vega,F.M., Moore,B.L., Gonzalez-Porta,M., Eberle,M.A., Tezak,Z., Lababidi,S., *et al.* (2019) Best practices for benchmarking germline small-variant calls in human genomes. *Nat. Biotechnol.*, **37**,

555–560.
41. GitHub - Illumina/hap.py: Haplotype VCF comparison tools.
42. Cruz-Tapias,P., Castiblanco,J. and Anaya,J.-M. (2013) Major histocompatibility complex: Antigen processing and presentation. In *Autoimmunity: From Bench to Bedside [Internet]*. El Rosario University Press.
43. Logsdon,G.A., Vollger,M.R., Hsieh,P., Mao,Y., Liskovykh,M.A., Koren,S., Nurk,S., Mercuri,L., Dishuck,P.C., Rhie,A., *et al.* (2021) The structure, function and evolution of a complete human chromosome 8. *Nature*, **593**, 101–107.
44. Martin,J., Han,C., Gordon,L.A., Terry,A., Prabhakar,S., She,X., Xie,G., Hellsten,U., Chan,Y.M., Altherr,M., *et al.* (2004) The sequence and analysis of duplication-rich human chromosome 16. *Nature*, **432**, 988–994.
45. Wang,J., Raskin,L., Samuels,D.C., Shyr,Y. and Guo,Y. (2015) Genome measures used for quality control are dependent on gene function and ancestry. *Bioinformatics*, **31**, 318–323.
46. Roden,D.M., McLeod,H.L., Relling,M.V., Williams,M.S., Mensah,G.A., Peterson,J.F. and Van Driest,S.L. (2019) Pharmacogenomics. *The Lancet*, **394**, 521–532.
47. Nurk,S., Koren,S., Rhie,A., Rautiainen,M., Bzikadze,A.V., Mikheenko,A., Vollger,M.R., Altemose,N., Uralsky,L., Gershman,A., *et al.* (2022) The complete sequence of a human genome. *Science*, **376**, 44–53.