Interactive visualization and interpretation of pangenome graphs by linear-reference-based coordinate projection and annotation integration

Zepu Miao and Jia-Xing Yue

Genome Res. published online January 13, 2025
Access the most recent version at doi:10.1101/gr.279461.124

Running title: Interpretative visualization of pangenome graphs

Interactive visualization and interpretation of pangenome graphs by linear-reference-based coordinate projection and annotation integration

Zepu Miao[1], Jia-Xing Yue[1,*]

[1]State Key Laboratory of Oncology in South China, Guangdong Key Laboratory of Nasopharyngeal Carcinoma Diagnosis and Therapy, Guangdong Provincial Clinical Research Center for Cancer, Sun Yat-sen University Cancer Center, Guangzhou 510060, China

*Corresponding author

E-mail: yuejiaxing@gmail.com (J.-X. Y.)

E-mail and ORCID for all authors:

Zepu Miao > zpmiao@sysucc.org.cn, 0000-0002-9734-682X

Jia-Xing Yue > yuejiaxing@gmail.com, 0000-0002-2122-9221

Abstract

With the increasing availability of high-quality genome assemblies, pangenome graphs emerged as a new paradigm in the genomics field for identifying, encoding, and presenting genomic variation at both population and species levels. However, it remains challenging to truly dissect and interpret pangenome graphs via biologically informative visualization. To facilitate better exploration and understanding of pangenome graphs towards novel biological insights, here we present a web-based interactive Visualization and interpretation framework for linear-Reference-projected Pangenome Graphs (VRPG). VRPG provides efficient and intuitive supports for exploring and annotating pangenome graphs along a linear-genome-based coordinate system (e.g., that of a primary linear reference genome). Moreover, VRPG offers many unique features such as in-graph path highlighting for graph-constituent input assemblies, copy number characterization for graph-embedding nodes, graph-based mapping for query sequences, all of which are highly valuable for researchers working with pangenome graphs. Additionally, VRPG enables side-by-side visualization between the graph-based pangenome representation and the conventional primary-linear-reference-genome-based feature annotations, therefore seamlessly bridging the graph and linear genomic contexts. To further demonstrate its functionality and scalability, we applied VRPG to the cutting-edge yeast and human reference pangenome graphs derived from hundreds of high-quality genome assemblies via a dedicated web portal and examined their local genome diversity in the graph contexts.

Contact: yuejiaxing@gmail.com

Introduction

Long-read sequencing technology has become the go-to-choice for most genome sequencing projects in recent years, empowering the production of chromosome-level telomere-to-telomere (T2T) genome assemblies for diverse organisms, including human (Yue et al. 2017; Jiao and Schneeberger 2020; Nurk et al. 2022). With such T2T reference assembly panel continuously expanding at both population and species levels, researchers began to use pangenome graphs to better represent the population- and species-wide genomic variation landscapes in a sequence-resolved manner (Eizenga et al. 2020). Compared with the conventional linear reference genome, pangenome graphs offer enhanced power and accuracy in read mapping and variant identification, especially in the presence of sequence polymorphisms and structural variants (SVs) (Paten et al. 2017). Therefore, a species-representative pangenome graph is expected to shed novel insights into the interpretation of the genotype-to-phenotype association and the discovery of missing heritability.

While a number of tools have been developed to build pangenome graphs based on genome alignments, Minigraph (Li et al. 2020), Minigraph-Cactus (Hickey et al. 2023), and PGGB (Garrison et al. 2024) are among the most popular ones. Minigraph is designed for efficiently constructing a primary linear-reference-based pangenome graph with large variants (e.g., SVs) compactly encoded in the reference Graphical Fragment Assembly (rGFA) format. The rGFA format records the source information of each graph-embedding node relative to the input linear genomes and allows for traversing the pangenome graph along a stable coordinate system. This unique feature makes the pangenome graph built by Minigraph a natural and intuitive extension to the conventional linear reference genome, although certain level of bias could be introduced during the graph building process regarding the choices of the primary linear reference and the input genome order (Garrison and Guarracino 2023). Minigraph

preferentially considers large variants (e.g., SVs) while being less discriminative for small variants such as single-nucleotide-variants (SNVs) and small insertion/deletions (indels) during graph construction, therefore a pangenome graph constructed by Minigraph is not a strictly lossless representation of the full genomic variation carried by input genomes at the per-base level. As an improvement, Minigraph-Cactus was proposed as an extended solution that combines the compactness and efficiency of Minigraph as well as the base-level sensitivity and accuracy of Cactus. Pangenome graphs constructed by Minigraph-Cactus can effectively represent both small and large genomic variants while still preserving a trackable coordinate system derived from the primary linear reference genome. Finally, PGGB adopted an alternative approach for pangenome graph construction based on all-against-all genome alignments, which is theoretically unbiased but computationally more intensive. Also, PGGB will clip and collapse the highly similar genomic segments of the input genomes and therefore disrupt the linearity of input genome coordinates, making it less straightforward in referring back to the original input genomes.

An intuitive visualization of pangenome graphs can greatly assist the exploration and understanding of the global and local genomic variation in their graph representations. To date, several tools have been developed for visualizing pangenome graphs. Among them, Bandage (Wick et al. 2015) , GFAviz (Gonnella et al. 2019), and gfaestus (Fischer 2024) focus more on large-scale graph topology, while SequenceTubeMap (Beyer et al. 2019), MoMI-G (Yokoyama et al. 2019), and PGR-TK (Chin et al. 2023) are more suitable for picturing fine-scale sequence variation. In addition, ODGI (Guarracino et al. 2022) showed improved performance on large-scale pangenome graphs with extended visualization function for binned and linearized 1-dimensional local graph structure rendering. Another tool, PanGraphViewer (Yuan et al. 2023), while still being built for graph topology visualization, enables a coordinate-based graph

querying and can be used for genomic variant examination when a VCF file is provided. While these tools are quite useful in specific use scenarios, they almost exclusively focus on graph per se, making it still challenging for general researchers to associate a pangenome graph with conventional linear genome assemblies and their feature annotations. Moreover, many of these existing tools fall short in scalability, lacking the capability of dynamically visualizing full-scale pangenome graphs on the fly. Therefore, a novel pangenome graph visualization framework is needed to address these shortcomings and to better bridge the linear-based and graph-based genomic representations, which will make pangenome graphs more accessible to a broader community.

Results

General design

In this study, we developed a web-based interactive Visualization and interpretation framework for linear-Reference-projected Pangenome Graphs (VRPG) with enhanced functionality and scalability. Functionality-wise, VRPG can present pangenome graphs along a stable linear coordinate system (e.g., that of a primary linear reference genome), therefore enabling browsing, querying, labeling, and highlighting pangenome graphs in a highly intuitive manner. For doing so, VRPG natively supports the rGFA-formatted pangenome graphs built by Minigraph while also shipped with an auxiliary command-line module (gfa2view) to provide extended compatibility to pangenome graphs in GFAv1 format (e.g., those built from Minigraph-Cactus and PGGB). In addition to visualizing the pangenome graph itself, VRPG allows for user-defined annotation tracks alongside, which unifies the pangenome graph with various annotation data types under the same primary-linear-reference-based coordinate

system. Regarding scalability, VRPG is algorithmically optimized to be capable of rapidly navigating, querying, and rendering large-scale pangenome graphs built upon hundreds of input genome assemblies. Multiple layout simplification options are further implemented to make it amenable for pangenome graphs with high complexity. Taken together, VRPG offers a novel and powerful way of visualizing pangenome graphs with unique strength in both functionality and scalability.

User interface and feature highlights

VRPG mainly operates via a web browser, in which users can easily navigate and interact with the rendered pangenome graph (Figure 1). The user interface of VRPG consists of six panels: (1) the control panel, (2) the graph panel, (3) the genome coordinates and gene annotation panel, (4) the additional feature annotation panel, (5) the node information panel, (6) the path information panel. Together, these panels offer a unified and engaging user experience for exploring pangenome graphs in a highly interactive and informative manner.

In the control panel, users can specify their interested pangenome graphs and query regions based on the predefined primary linear reference genome coordinates. VRPG can rapidly extract the corresponding subgraph accordingly and render it in five optional layouts (i.e., “Ultra expanded” (default), “Expanded”, “Squeezed”, “Hierarchical expanded”, “Hierarchical squeezed”), three graph simplification strategies (i.e., “Nonref nodes” (default), “All nodes”, and “None”), and an additional simplification parameter (i.e., “Min bubble size”, which is approximately the minimal cumulative node length within a bubble). These simplification options can be very useful when visualizing pangenome graphs built by Minigraph-Cactus or PGGB. In such pangenome graphs, base-level small variants are encoded as individual nodes, therefore resulting in a topology that is too complex to display unless properly simplified.

While the layout and simplification options described above already offer considerable flexibility for users to find the best rendering results, users can also manually drag any node in the displayed graph for layout fine-tuning when needed. Additional functions such as assembly-to-graph path highlighting (the "Path highlighting" button) and sequence-to-graph mapping (the "Sequence-to-graph mapping" button) can be further used for more advanced graph exploration.

In the graph panel, graph nodes corresponding to the predefined primary linear reference genome are always scaled according to their actual size and typically shown along the center line, while nodes that represent genomic variants relative to the primary linear reference are shown alongside, which can be optionally scaled. Both the displayed nodes and edges are clickable. For nodes, the selected node will be thickened upon clicking, and its corresponding information will be reported in the node information panel. As for edges, the selected edge will be colored in red on the first click and its color will revert to black on the second click. When the assembly-to-graph path highlighting feature is enabled, the traversing paths of the selected query assembly(-ies) will be marked in the graph, with all matched nodes and edges highlighted in red. The arrows on the matched edges reflect the traversing direction of the highlighted linear genome assemblies in the pangenome graph.

In the genome coordinates and gene annotation panel, a primary-linear-reference-based coordinate track is shown to provide positional reference to the pangenome graph rendered above. Together with this coordinate track, gene and mRNA annotation tracks are further depicted to provide more functional contexts. Users can find out the name of the displayed genes and mRNAs by mouse clicking (for genes) or hovering (for mRNAs).

In the additional feature annotation panel, users can specify one or more primary-linear-reference-based annotation features and visualize them together with the pangenome graph in a side-by-side and highly synchronized manner. Both qualitative and quantitative annotation features are supported here in the BED format. In this way, users can essentially display any annotation features (e.g., GC content, centromere, telomere, conserved elements, regulatory elements, genomic variants) and conveniently explore their physical association with the topology of pangenome graphs.

In the node information panel, a detailed summary report on node-associated information will be automatically generated upon users' node selection in the graph. The reported information includes the assembly origin of the selected node, genes covered by this node, and estimated copy numbers of the corresponding genomic sequence across different graph-constituent assemblies.

In the path information panel, when users perform assembly-to-graph path highlighting (if only one assembly is selected) or sequence-to-graph mapping (currently only support pangenome graphs built by Minigraph), the returned results will be shown here describing the matched query genome coordinates, graph traversing path (denoted by directionally chained node IDs), and concise idiosyncratic gapped alignment report (CIGAR) string. This feature offers the precise measurement on how well the query sequence (specified for highlighting or mapping) matched with the pangenome graph, which is of great value for graph-based comparative genomic analysis but remains unavailable in any other pangenome graph visualization tools developed so far.

Genomic variant representation in VRPG

One major motivation of developing VRPG is to provide an intuitive way of associating the conventional linear-assembly-based genomic variants to their graph representations. Now it comes in handy with VRPG's assembly-to-graph path highlighting feature. In Figure 2, we exemplified how different types of genomic variants from the query genome relative to the primary linear reference are typically depicted by VRPG in the context of a pangenome graph (Figure 2). Since VRPG typically shows nodes representing the primary linear reference assembly along the center line, any highlighted query genome path that departs from this center line implies the occurrence of genomic variant(s). Moreover, since VRPG's highlighting function also depicts the traversing direction of each path from one node to another, it is straightforward to determine the genomic orientation of the variant(s) observed (e.g., in the case of an inversion). In addition, the in-node paths will also be thickened in a two-state manner (thin vs. thick) if the highlighted genome assembly traverses the corresponding node(s) more than once, which implies the occurrence of duplications. Aside from these graphical indications, the detailed traversing path of the highlighted assembly are also reported in the Path information panel of VRPG, which provides additional help for variant interpretation. The traversing path denotation used by VRPG follows the Graph Alignment Format (GAF) specification.

To demonstrate the application of VRPG in real world examples, we set up a dedicated webserver (https://www.evomicslab.org/app/vrpg/) to visualize a reference pangenome graph derived from 163 budding yeast genome assemblies. Here we employed Minigraph to construct this pangenome graph using the *Saccharomyces cerevisiae* reference assembly panel (ScRAP) that we recently assembled (O'Donnell et al. 2023; Miao et al. 2024). This yeast pangenome graph consists of 37,062 nodes and 52,756 edges, with a total length of 27,190,479 bp.

The flip/flop inversion

Based on this yeast pangenome graph, we used VRPG to visualize a famous flip/flop inversion region on the Chromosome XIV (ChrXIV) of the *S. cerevisiae* genome. This flip/flop inversion region is flanked by two 4.2-kb inverted repeat (IR) regions and remains polymorphic in *S. cerevisiae* and its sister species in the genus *Saccharomyces* (Salzberg et al. 2022). For example, within *S. cerevisiae*, our previous study revealed that the genome of the Sake strain Y12 shares conserved synteny with the *S. cerevisiae* reference genome (SGDref) within this region whereas the genome of the North American strain YPS128 shows an inversion instead (Yue et al. 2017) (Figure 3A). In accordance with this prior knowledge, by enabling the assembly path highlight feature of VRPG, we recaptured such polymorphic inversion in the pangenome graph (Figure 3B-E). As illustrated by VRPG, both SGDref and Y12 revealed a simple linear assembly-to-graph path through the nodes along the central line, suggesting their primary-linear-reference-like sequence structure in this region (Figure 3B, D). In contrast, YPS128 shows an S-shaped path in the flip/flop region, which implies the existence of the inversion (Figure 3E).

The *DOG2* deletion

The yeast paralog gene pairs *DOG1* and *DOG2* encode 2-deoxyglucose-6-phosphate phosphatase involved in glucose metabolism. They are homologous to the human *PUDP* (previously known as *HDHD1*) gene, an anti-cancer treatment target for intervene the glycolytic metabolism of tumor cells (Defenouillère et al. 2019). Previously we have identified a polymorphic deletion for the *DOG2* gene in the comparison of seven representative yeast strains using their T2T genome assemblies (Yue et al. 2017). For example, this gene is present in strains like S288C and W303, but absent in other strains such as SK1 and Y12 (Figure 4A).

Here we used VRPG to visualize this gene presence/absence variation in the context of a pangenome graph (Figure 4B-F). With the SGDref (S288C) as the primary linear reference, the path taken through the graph by the linear assembly W303 is simple, suggesting its reference-like sequence structure in this region. In contrast, the paths taken by SK1 and Y12 clearly depart from the SGDref path by bypassing the node representing the *DOG2* gene region.

Application demonstration 2: visualizing pangenome graphs derived from 90 human genome assemblies.

As a further proof for VRPG's versatility and scalability, we retrieved the human pangenome graphs generated by the Human Pangenome Reference Consortium (HPRC) based on 90 human genome assemblies (Liao et al. 2023) and visualized them via our VRPG demonstration webserver (https://www.evomicslab.org/app/vrpg/). Starting with the same input genome set, HPRC built three human pangenome graphs with Minigraph, Minigraph-Cactus, and PGGB respectively. The HPRC human pangenome graph built by Minigraph consists of 391,950 nodes and 566,204 edges, with a total length of 3,198,196,033 bp. In comparison, the total node and edge numbers of graphs built by Minigraph-Cactus (81,415,956 nodes and 112,955,105 edges) and PGGB (110,884,673 nodes and 154,756,169 edges) are substantially larger since small variants such as SNV and indels are fully considered by Minigraph-Cactus and PGGB during their graph construction whereas Minigraph predominantly considers larger variants such as SVs.

The *DSCAM* intronic inversion

For the demonstration, we used VRPG to visualize an inversion located between the exon 32 and exon 33 of the *DSCAM* gene (Audano et al. 2019) (Figure 5). This inversion is mediated by two inverted 6.0-kb L1 elements (Figure 5A), forming a structure much like the yeast

ChrXIV flip/flop inversion described above. *DSCAM* gene belongs to the immunoglobulin superfamily of cell adhesion molecules (Ig-CAMs) and functions in central and peripheral nervous system development. The overexpression of *DSCAM* promotes the development of Down syndrome (DS) and congenital heart disease (DSCHD) (Grossman et al. 2011; Liu et al. 2023). This *DSCAM* inversion is segregated between the human reference assembly GRCh38 and the telomere-to-telomere (T2T) assembly CHM13 (Figure 5B). Using VRPG's assembly path highlighting feature, we compared the paths of GRCh38 and CHM13 in HPRC human pangenome graphs built from Minigraph, Minigraph-Cactus, and PGGB (Figure 5C-E). In all graphs, the GRCh38 shows a simple linear path as expected. The CHM13 T2T assembly, on the other hand, shows the characteristic S-shaped pattern in both Minigraph and Minigraph-Cactus pangenome graphs for the inverted region. In the PGGB graph, the pattern is more complex, likely due to the sequence collapsing procedure during PGGB's graph construction. Nevertheless, clear distinctions between the GRCh38 and CHM13 paths can still be found, as CHM13 shows a unique protruding loop in the inverted region. Moreover, given that the pangenome graphs built from Minigraph-Cactus and PGGB tend to be much more complex in topology, here we used this case to show the effectiveness of VRPG's layout simplification feature. When no layout simplification was applied, we can see signals of many genomic variants (indicated with small triangles in Figure 6) in the Minigraph-Cactus and PGGB graphs. When we enabled the nonreference node simplification, most such signals disappeared as they represent small variants (i.e., <50 bp), leaving the large variants (e.g., our interested inversion) more noticeable even without zooming-in.

The *CR1* intragenic deletion

The *CR1* gene (also known as *CD35*) is a complement activation receptor gene that is strongly associated with the Alzheimer disease (Lambert et al. 2009; Brouwers et al. 2012). This gene

is structurally polymorphic in human populations, bearing an intragenic duplication region whose copy number correlates with the risk of Alzheimer disease (Kucukkilic et al. 2018). An 18.6-kb intragenic deletion has been reported for *CR1* in a recent comparison between the human GRCh38 reference genome and the CHM13 T2T genome, with the latter one hosting the shorter version (Yang et al. 2023). Here we took a closer examination for this deletion in the context of its nearby sequence homology and confirmed that this deletion locates within the previously reported intragenic duplication region (Figure 6A-B). Next, we used VRPG to visualize this *CR1* intragenic deletion based on the human pangenome graphs built by Minigraph, Minigraph-Cactus, and PGGB respectively (6C-E). By comparing the highlighted paths of GRCh38 and CHM13, this deletion can be clearly identified in both Minigraph and Minigraph-Cactus pangenome graphs. As for the PGGB graph, VRPG also depicted notable path differences between GRCh38 and CHM13 (note those extra red paths highlighted for CHM13), although the PGGB graph is topologically too complex to reflect the deletion in an intuitive way.

Functionality and performance comparison with other pangenome graph visualization tools

As briefly described in the introduction, there have been several tools developed for visualizing pangenome graphs. A representative list of these tools may include Bandage, GfaViz, gfaestus, SequenceTubeMap, MoMI-G, ODGI, PGR-TK, and PanGraphViewer. Here we summarized their basic information and design features in a comparison table (Table 1). In comparison, VRPG shines in its all-around input compatibility, web-based dynamic display, and versatile functionalities. Especially, being able to carry out cross-examination between the pangenome graph and linear assemblies is critical for understanding and utilizing a pangenome graph, which is currently best supported in VRPG. Moreover, although several tools can take

additional genome annotation files as the inputs, how they can actually utilize these files vary substantially. For example, Bandage can use such annotation files for region-specific graph highlighting while PanGraphViewer uses them for subgraph extraction instead. VRPG is the only tool that enables a highly interactive and synchronized visualization between a pangenome graph and rich genomic feature annotations, helping to place the somewhat abstract pangenome graph into a more intuitivend biologically informative context.

Aside from the comparison in general designs and functionalities, we also benchmarked the computational performance of VRPG against Bandage, GfaViz, and PanGraphViewer for the yeast and human pangenome graph visualization (Table 2). All of these four tools support GFAv1/rGFA-formatted pangenome graphs, which makes them more comparable than other tools. First, we tested if these tools could parse (no need for rendering) the full-scale yeast and human pangenome graphs used in this study. The Chr22 subset of the full-scale human graphs were also used for additional comparison. Note that by design VRPG performs format transformation (when the input graph is in GFAv1 format) and indexing in advance, which are its heavy-lifting steps and consumes more computational resources. Considering that these steps only need to be executed once and it can be fully prepared before the actual graph parsing and visualization, these steps were not included in our benchmarking analyses (See Methods for an estimated resource consumption based on HPRC human pangenome graphs). Mechanistically, VRPG, Bandage, and PanGraphViewer all adopted a two-step strategy for graph visualization, with the parsing and rendering steps fully decoupled. In contrast, GfaViz parses and renders the graph in a single step, thus consuming substantially more running time and memory. As a result, GfaViz was not able to parse the full-scale human graphs, neither for their Chromosome 22 (Chr22) subsets, as it quickly hit the memory cap of our testing machine (16 GB). Bandage also struggled with the full-scale Minigraph-Cactus and PGGB graphs due

to memory limitation. PanGraphViewer has better memory management but appears less robust when parsing graphs built by Minigraph-Cactus and PGGB. In comparison, VRPG showed highly stable and robust performance when parsing all tested pangenome graphs once the corresponding graphs have been properly converted and indexed in advance.

Next, we compared the computational resource consumption of graph rendering for specific genomic regions. Four regions were selected for this test, two from the yeast graph and two from the human graphs. As for the full-graph, we found GfaViz is more sensitive to graph complexity as it failed to render the Minigraph-Cactus-based and PGGB-based subgraph for the 500-kb testing region (Chr22:20,000,001-20,500,000) in human. Also, PanGraphViewer seems lacking the support for visualizing PGGB-based subgraph as observed in our full-graph test. Both Bandage and VRPG showed robust performance in subgraph rendering, with VRPG further shined in rendering speed, thanks to its block index design (see Methods).

Discussion

As genome sequencing technologies keep progressing towards longer read length and higher per-base accuracy, generating reference-quality genomes at a population scale is becoming increasingly affordable (De Coster et al. 2021). Along with this trend, researchers began to develop new computational frameworks to accommodate and analyze such growing collection of high-quality genomes with better efficiency. The pangenome graph serves as a compact yet extensible data structure that describes population-level genetic diversity at both base and structure levels via a graph representation (Eizenga et al. 2020). This is a highly active research field with new algorithms and applications rolling out very quickly. One can envision that many genomics analyses that currently rely on conventional linear reference genomes will be eventually carried out based on pangenome graphs in future. A better understanding of how

different layouts and topologies of pangenome graphs correspond to the actual population-wide genomic variation is of critical importance for correctly interpreting pangenome-graph-based analysis results.

In this study, we developed VRPG, a web-based interactive framework for pangenome graph visualization and interpretation with full scalability, capable of rendering complex pangenome graphs derived from hundreds of genome assemblies. In addition to intuitive graphical visualization, VRPG also provides native supports for integrating conventional linear-genome-based feature annotations, enabling seamless examination of genomic variation in both graph-based and linear-based contexts. As further demonstrated with real world examples of yeast and human pangenome graphs as well as with benchmarking comparison with other tools, VRPG shines with its unique advantages in both functionality and scalability, making it a highly compelling choice for interactive and informative pangenome graph visualization.

Although efforts have been taken, there are also some limitations for VRPG in its current form. For example, some of its display layouts can be further improved for better clarity, especially when the graph topology become complex. Its sequence-to-graph mapping feature currently only works for the pangenome graphs built by Minigraph, whereas similar functionality is yet to be developed for graphs built by Minigraph-Cactus and PGGB. In the same vein, a broader graph compatibility beyond the rGFA and GFAv1 formats will be quite helpful for extending VRPG's application scope. Finally, the installation of VRPG has only been tested in Linux-based environment (e.g., CentOS and Ubuntu) so far. Additional installation compatibility with MacOS should be possible but yet to be explored. That said, given the web-based design of VRPG, users can access it remotely from any web-enabled platform via a centralized installation, which should help to alleviate this limitation. In the near future, with the help from

394 the ever-growing pangenome graph community, we anticipate VRPG to be further improved,
395 facilitating researchers from different fields to better explore the power of graph-based
396 pangenomics, especially in the age of large-scale long-read-based population sequencing
397 (O'Donnell et al. 2023; Weller et al. 2023; Rech et al. 2022; Lian et al. 2024; Gustafson et al.
398 2024).
399

400 # Methods

401 ## Hardware and software recommendations

402 VRPG is designed for a web-enabled desktop or computing server running the Linux operating
403 system. Regarding the hardware, the resource-intensive steps are graph2view transformation
404 (when input graph is GFAv1-formatted) and index building. For example, for the HPRC
405 Minigraph-Cactus graph, these two steps together took ~4 hours with 10 CPU threads and 16-
406 GB memory on a computing server equipped with Intel(R) Xeon(R) Gold 6248R CPU (@
407 3.00GHz). Therefore, we recommend executing these two steps on a computing server in
408 advance when the input pangenome graph is large and complex. But once the format
409 transformation and index building are processed, very little resource is further needed to access
410 and render the indexed graph within the VRPG framework. Even with the HPRC human
411 pangenome graphs, an ordinary desktop with 4 GB memory should be enough to render the
412 subgraph for any given region that is <=1 Mb. As for the software, VRPG has been tested with
413 both CentOS and Ubuntu Linux operating systems. A list of third-party software dependencies
414 for VRPG's installation, compilation, and execution is further provided (Supplemental Table
415 1).
416

Naming scheme of input assemblies

For VRPG, the naming scheme of input assemblies follows a relaxed version of the PanSN prefix naming specification (https://github.com/pangenome/PanSN-spec). The input assembly identifier should consist of three parts: sample tag, delimiter, and haplotype tag. The sample tag and haplotype tag can be any strings consist of letters and/or numbers, while the delimiter should be a user-specified character (option: --sep) that is not used in the assembly/contig names (e.g., # or :). The maximum number of assemblies allowed by VRPG is 65535.

Input pangenome graph specification

In a pangenome graph, a node represents a directed genomic segment while an edge represents the relative order and direction of the two inter-connected nodes. VRPG natively works with pangenome graph files in the rGFA format, which can be straightforwardly constructed using Minigraph (Li et al. 2020). Compared with pangenome graph encoded in other formats, the rGFA-formatted pangenome graph inherently tracks the origin of each node in terms of its location and direction from the input assembly, which comes handy when analyzing and interpreting the corresponding graph. In a rGFA-formatted graph, all nodes from the predefined primary linear reference are labeled as rank-0 (defined with the SR tag) while those incrementally added non-primary-linear-reference nodes derived from other graph-constituent assemblies are labeled with lower ranks (e.g., 1, 2, etc.). VRPG can readily use this information for organizing the displayed graph layouts during graph visualization. While the rGFA-graph does not store the assembly-to-graph path traversing information, such information can be straightforwardly recovered by aligning each graph-constituent assembly to the rGFA-graph using Minigraph. In addition to the rGFA-encoded pangenome graphs, VRPG also supports GFAv1-encoded pangenome graphs with a dedicated module (gfa2view) for format transformation.

GFAv1 format transformation for VRPG

GFAv1 is a widely used pangenome graph format. Unlike the rGFA-formatted pangenome graph, there is no predefined primary linear reference for establishing the coordinate system in GFAv1-encoded graphs. Therefore, to accommodate GFAv1-formated graph with VRPG, a user defined reference genome is required. When there are multiple copies of highly similar sequence segments (e.g., in the case of segmental or tandem duplication) in the same genome assembly, some pangenome graph builders such as PGGB tends to collapse them into a single node, which is more efficient in compressing genomic information but also breaks the linearity of the reference genome coordinate system. To work around this challenge, the gfa2view module of VRPG will traverse the reference path to find the nodes that are traversed for two or more times (i.e., the collapsed nodes) and insert the corresponding “shadow nodes” and their associated edges into the graph to restore the original reference coordinates for the given node. In this way, the primary linear reference genome rendered by VRPG still keeps linearity. The resulting graph with these added shadow nodes is recorded in gfa2view’s output file, which can be used for VRPG as the input file for visualization. Meanwhile, the predefined path (the P-line) and walk (the W-line) information in the GFA file will be extracted by VRPG for tracing the assembly-to-graph path of each graph-constituent assembly. The theoretical maximal node allowance in VRPG is 2147483647. In practice, this upper limit will be lower for PGGB graphs, given that newly introduced “shadow nodes” consume available node indices. There is no upper limit for edge counts in VRPG.

Node and edge rendering

Representative genomic segments are denoted by graph nodes and further illustrated as colored blocks in the view window. For nodes representing the primary linear reference genome

(predefined when building the reference pangenome graph), their block sizes are generally proportional to the actual size of the corresponding genomic segments. For nodes representing nonreference assemblies, their corresponding blocks can be optionally scaled. If the length of a node (denoted as "S" herein) is <2 kb, the node will be illustrated as a rectangular block. Otherwise, the node will be represented by a curved block where the corner count of the curved block is proportional to the multiples of 2-kb determined by S. The connections between different nodes are represented by graph edges and further illustrated as directed lines. For the cleanness and efficacy of rendering, additional graph simplification algorithms are further applied to trim off those peripheral nodes that are far away from the reference nodes (e.g., when traversing depths are greater than or equal to user defined search depth) as well as to optionally hide small nodes (e.g., those representing genomic variants that are <50 bp) in the graph. These features are especially helpful when visualizing genome graphs built by Minigraph-Cactus and PGGB where all base-level variants (e.g., SNVs and indels) are denoted as nodes. In practice, we recommend a maximal query region size of 1 Mb to restrain the total number of nodes and edges for rendering, so that the web browser will not get overwhelmed during graph rendering.

Efficient access to the graph

VRPG implements a block index system to enable quick access to the subgraph corresponding to any given region along the predefined linear reference genome. Briefly, the predefined linear reference genome is first subdivided into blocks with each block containing 2,000 reference nodes. For each block, a breadth-first search algorithm was implemented to find all edges and nonreference nodes associated with the reference nodes within the corresponding block and a block index corresponding to these edges and nodes was created. To reduce the search space and accelerate the index building process, by default, given a predefined primary linear reference chromosome, VRPG considers nodes that fall into the following two scenarios: 1)

nodes with at least one edge directly connecting to the focal chromosome, 2) nodes with edges that are neither directly connecting to the focal chromosome nor being private to any other chromosome. Based on the indexed nodes and edges, an ordered block index array for each assembly path was created. With such block index system, VRPG can efficiently locate the subgraph associated with the query region (either directly overlapping or connected by edges) for rendering. VRPG uses the '--xDep' parameter (specified via vrpg_preprocess.py or gfa2view) to set the hard maximal search depth allowance for graph traversing during block indexing. When setting this value large enough (default: 100), VRPG should be able to exhaustively retrieve all potentially relevant material for rendering. Note that another 'search depth' option (default: 10) is also provided to set the soft maximal search depth allowance for the final graph rendering, which is inherently constrained by the hard limit set by '--xDep'. In the case of a common node shared by multiple blocks or even chromosomes within the specified search depth allowance, such association will be recorded for all related blocks and chromosomes so that the node can always be accessed via its associated blocks and chromosomes.

Yeast reference pangenome graph construction

The yeast reference pangenome graph was constructed for this study. This graph was built upon 163 yeast genome assemblies from 142 strains, with some heterozygous/polyploid strains having both phased and collapsed assemblies. Briefly, we took the *Saccharomyces cerevisiae* reference genome (version: R64-1-1; denoted as SGDref) retrieved from *Saccharomyces* genome database (SGD; https://www.yeastgenome.org/) as well as 162 assemblies from our recently released *Saccharomyces cerevisiae* Reference Assembly Panel (ScRAP) (O'Donnell et al. 2023; Miao et al. 2024) to construct reference pangenome graph. Minigraph (Li et al. 2020) was used for this graph construction with the command 'minigraph -cxggs -l 5000'. With

the SGDref as the reference genome, we incrementally added those 162 ScRAP assemblies into the graph according to their phylogenetic distances to SGDref. The phylogenetic distances employed here were extracted from the phylogenetic tree of these 163 input genomes built upon their concatenated 1-to-1 orthologous gene matrix previously described (O'Donnell et al. 2023). Regarding the haplotype tag, for the yeast reference pangenome graph, we used "HP0" to denote haplotypes of haploid or homozygous diploid strains, while using "collapsed", "HP1", and "HP2" to denote collapsed, or the two phased haplotypes of heterozygous diploid strains. The constructed pangenome graph is provided in Supplemental Materials (Supplemental Data File) and has also been deposited at Zenodo (see Software availability).

Human reference pangenome graph construction

The HPRC constructed three human pangenome graphs using Minigraph, Minigraph-Cactus, and PGGB respectively based on 90 genome assemblies from 46 samples (Liao et al. 2023). In addition to the classic (GRCh38) and T2T (CHM13) human reference genome assemblies, the other 44 samples all have two fully phased genome assemblies. For the Minigraph graph, we used the version deposited by Dr. Heng Li at (https://zenodo.org/record/6983934) (Li 2022) with the download link: https://zenodo.org/records/6983934/files/GRCh38-90c.r518.gfa.gz. For the Minigraph-Cactus graph, we used the HPRC v1.1 version, accessible via the download link: https://s3-us-west-2.amazonaws.com/human-pangenomics/pangenomes/freeze/freeze1/minigraph-cactus/hprc-v1.1-mc-grch38/hprc-v1.1-mc-grch38.gfa.gz. For the PGGB graph, we used the HPRC v1.0 version, accessible via the download link: https://s3-us-west-2.amazonaws.com/human-pangenomics/pangenomes/freeze/freeze1/pggb/hprc-v1.0-pggb.gfa.gz. For these human reference pangenome graphs, the haplotype tag "0" was used to denote haplotypes of haploid samples or those where the haplotype is indeterminate due to mosaicism or collapsing, while

"mat" and "pat" were used to denote the maternal and paternal haplotypes of the phased diploid samples. Note that due to the disk space limitation of our web server, we opted for not storing the node sequences for the human pangenome graphs on our demonstration site. Therefore, when clicking on a node in the demonstrated human pangenome graphs, the node sequence will not be reported in the node information panel, which is not the case for the demonstrated yeast pangenome graph.

Annotation track preparation

For the annotation tracks displayed with the yeast pangenome graph, the SGDref's gene, centromere, and telomere annotations were extracted from the NCBI RefSeq GFF3 file (https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/146/045/GCF_000146045.2_R64/GCF_000146045.2_R64_genomic.gff.gz). The subtelomere annotation was extracted from the GFF3 file retrieved from the ScRAPdb database (https://www.evomicslab.org/db/ScRAPdb/)(Miao et al. 2024). The GC% was calculated with 250-bp nonoverlap windows using the profile_GC_sliding_window.pl script (options: -w 250 -s 250) from RecombineX (Li et al. 2022). The conserved elements identified based on the whole-genome alignment of seven budding yeast species (phastConsElements7way) was retrieved from the Table Browser tool provided by the UCSC Genome Browser (Raney et al. 2024).

For the annotation tracks displayed with the human pangenome graph, the GRCh38-based gene annotation was extracted from the NCBI RefSeq GFF3 file (https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/001/405/GCF_000001405.40_GRCh38.p14/GCF_000001405.40_GRCh38.p14_genomic.gff.gz). The centromeres, Genome Aggregation Database-SVs (gnomAD_SV), and conserved elements identified based on the whole-genome alignment of 100 vertebrate species (phastConsElements100way) were retrieved from UCSC Genome Browser's Table Brower as described above. For the "High

567 priority clinical genes" track, a previously curated gene set with high-priority clinical
568 significance (merged set: 857 genes) was obtained from the literature (Wagner et al. 2022) and
569 their corresponding genome coordinates were extracted based on the NCBI gene annotation
570 GFF3 file. The GC% was calculated with 2500-bp nonoverlap windows using the
571 aforementioned profile_GC_sliding_window.pl script (options: -w 2500 -s 2500).
572

573 ## Software implementation and web deployment

574 The backend of VRPG is implemented in C++ and Python3 based on the Django framework
575 (https://www.djangoproject.com/), while D3.js (https://github.com/d3/d3) was employed at the
576 frontend for data visualization. Bootstrap and jQuery were further used for interactive query
577 and rendering. The demonstration webserver is deployed via an Alibaba Simple Application
578 Server with 2 CPU, 4 Gb RAM, and 280 Gb ESSD data storage. VRPG (v0.1.5) was used for
579 this web demonstration.
580

581 ## Linear genome comparison and visualization for the demonstrated examples

582 The genomic regions of the demonstrated loci (the ChrXIV flip/flop inversion and *DOG2* for
583 yeast as well *DSCAM* and *CR1* for human) were extracted from the corresponding linear
584 genome assemblies and subsequently compared and plotted with EasyFig (GitHub commit:
585 639daec) (Sullivan et al. 2011).
586

587 ## Pangenome graph visualization benchmarking

588 We systematically evaluated the graph parsing and rendering performance of VRPG (v0.1.5),
589 Bandage (v0.9.0), GfaViz (v1.0) and panGraphViewer (v1.0.2; desktop version). For graph

parsing, the full-scale yeast and human pangenome graphs as previously described were used. Additionally, the human Chr22 subgraphs from all three full-scale human pangenome graphs were further extracted by GFAtools (v0.5-r292-dirty; for pangenome graphs built from Minigraph and Minigraph-Cactus) and ODGI (v0.9.0-0-g1895f496; for pangenome graphs built from PGGB). For graph rendering, while VRPG and PanGraphViewer can directly access the corresponding subgraph based on query coordinates, Bandage and GfaViz can only extract subgraph based on queried node IDs. Therefore, to better gauge rendering time and peak memory usage, we prepared the subgraphs for our benchmarked regions with GFAtools and ODGI as described above. The genomic regions for such subgraph extraction include the SGDref-based ChrIV:1-50000 and ChrIV:1-500000 regions for yeast, the GRCh38-based Chr22, Chr22:2000001-2100000, and Chr22:20000001-20500000 regions for human.

All benchmark tests were performed using a desktop computer equipped with Intel Core i5-9400 CPU, 16 GB RAM, and 1TB hard disk and running with the Ubuntu 24.04.1 LTS operating system. To track the peak memory consumption, the GNU time (v1.9; option: -v; https://www.gnu.org/software/time/) was used for the tests of Bandage, GfaViz, and panGraphViewer. For VRPG, given its web-based nature, the Chrome (v130.0.6723.91)'s task manager function was used for assessing the memory consumption. To track the graph parsing and rendering time, the stopwatch function of iPhone's Clock app (v1.1) was used for Bandage. For GfaViz, and panGraphViewer, their the natively reported graph parsing and rendering time was used. For VRPG, the Chrome's developer tools function was used to calculate browser's response time.

Software availability

VRPG is free for use under the MIT license, with the source code available in both Supplemental Code File and GitHub (https://github.com/codeatcg/VRPG). The real-case demonstration of VRPG on yeast and human pangenome graphs is hosted at https://www.evomicslab.org/app/vrpg/. The 163-yeast-genome-based pangenome graph constructed for this study is available in both Supplemental Data File as well as at Zenodo (https://zenodo.org/records/13968346/files/ScRAP_v20230121_163asm.minigraph.gfa.gz).

Acknowledgements

We thank the valuable comments and suggestions from the three anonymous reviewers, which helped to significantly improve the quality of this manuscript and the associated software. We thank the technical support from Huihui Li and Ludong Yang for software testing and benchmarking. We are grateful to Jing Li for critically reading the manuscript.

Funding

This work is supported by National Natural Science Foundation of China (32470663 and 32070592 to J.-X. Y.), Guangdong Pearl River Talents Program (2019QN01Y183 to J.-X. Y.), Fundamental Research Funds for the Central Universities of Sun Yat-sen University (24qnpy293 to J.-X. Y.), and Young Talents Program of Sun Yat-sen University Cancer Center (YTP-SYSUCC-0042 to J.-X. Y.). The funding agencies have not played any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

J.-X.Y. conceived the study. M.Z. developed the software. M.Z. and J.-X.Y. analyzed the results. J.-X.Y. and M.Z. wrote the paper. All authors reviewed and contributed to the final version of the paper. Correspondence to J.-X.Y.

Competing interests

The authors declare no competing interests.

References

Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AE, Dougherty ML, Nelson BJ, Shah A, Dutcher SK, et al. 2019. Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell* **176**: 663-675.e19.

Beyer W, Novak AM, Hickey G, Chan J, Tan V, Paten B, Zerbino DR. 2019. Sequence tube maps: making graph genomes intuitive to commuters. *Bioinformatics* **35**: 5318–5320.

Brouwers N, Van Cauwenberghe C, Engelborghs S, Lambert J-C, Bettens K, Le Bastard N, Pasquier F, Montoya AG, Peeters K, Mattheijssens M, et al. 2012. Alzheimer risk associated with a copy number variation in the complement receptor 1 increasing C3b/C4b binding sites. *Mol Psychiatry* **17**: 223–233.

Chin C-S, Behera S, Khalak A, Sedlazeck FJ, Sudmant PH, Wagner J, Zook JM. 2023. Multiscale analysis of pangenomes enables improved representation of genomic diversity for repetitive and clinically relevant genes. *Nat Methods* **20**: 1213–1221.

De Coster W, Weissensteiner MH, Sedlazeck FJ. 2021. Towards population-scale long-read sequencing. *Nat Rev Genet* **22**: 572–587.

Defenouillère Q, Verraes A, Laussel C, Friedrich A, Schacherer J, Léon S. 2019. The induction of HAD-like phosphatases by multiple signaling pathways confers resistance to the metabolic inhibitor 2-deoxyglucose. *Science Signaling* **12**: eaaw8000.

Eizenga JM, Novak AM, Sibbesen JA, Heumos S, Ghaffaari A, Hickey G, Chang X, Seaman JD, Rounthwaite R, Ebler J, et al. 2020. Pangenome Graphs. *Annual Review of Genomics and Human Genetics* **21**: 139–162.

Fischer C. 2024. chfi/gfaestus. https://github.com/chfi/gfaestus (Accessed July 10, 2024).

Garrison E, Guarracino A. 2023. Unbiased pangenome graphs. *Bioinformatics* **39**: btac743.

Garrison E, Guarracino A, Heumos S, Villani F, Bao Z, Tattini L, Hagmann J, Vorbrugg S, Marco-Sola S, Kubica C, et al. 2024. Building pangenome graphs. *Nat Methods* 1–5.

Gonnella G, Niehus N, Kurtz S. 2019. GfaViz: flexible and interactive visualization of GFA sequence graphs. *Bioinformatics* **35**: 2853–2855.

Grossman TR, Gamliel A, Wessells RJ, Taghli-Lamallem O, Jepsen K, Ocorr K, Korenberg JR, Peterson KL, Rosenfeld MG, Bodmer R, et al. 2011. Over-Expression of DSCAM and COL6A2 Cooperatively Generates Congenital Heart Defects. *PLOS Genetics* **7**: e1002344.

Guarracino A, Heumos S, Nahnsen S, Prins P, Garrison E. 2022. ODGI: understanding pangenome graphs. *Bioinformatics* **38**: 3319–3326.

Gustafson JA, Gibson SB, Damaraju N, Zalusky MPG, Hoekzema K, Twesigomwe D, Yang L, Snead AA, Richmond PA, Coster WD, et al. 2024. High-coverage nanopore sequencing of samples from the 1000 Genomes Project to build a comprehensive catalog of human genetic variation. *Genome Res* **34**: 2061–2073.

682 Hickey G, Monlong J, Ebler J, Novak AM, Eizenga JM, Gao Y, Marschall T, Li H, Paten B. 2023.
683 Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nat*
684 *Biotechnol* **42**: 663–673.

685 Jiao W-B, Schneeberger K. 2020. Chromosome-level assemblies of multiple Arabidopsis
686 genomes reveal hotspots of rearrangements with altered evolutionary dynamics. *Nat*
687 *Commun* **11**: 989.

688 Kucukkilic E, Brookes K, Barber I, Guetta-Baranes T, Morgan K, Hollox EJ, ARUK Consortium.
689 2018. Complement receptor 1 gene (CR1) intragenic duplication and risk of
690 Alzheimer's disease. *Hum Genet* **137**: 305–314.

691 Lambert J-C, Heath S, Even G, Campion D, Sleegers K, Hiltunen M, Combarros O, Zelenika D,
692 Bullido MJ, Tavernier B, et al. 2009. Genome-wide association study identifies
693 variants at CLU and CR1 associated with Alzheimer's disease. *Nat Genet* **41**: 1094–
694 1099.

695 Li H. 2022. Minigraph pangenome graphs for HPRC year-1 samples.
696 https://zenodo.org/record/6983934 (Accessed January 20, 2023).

697 Li H, Feng X, Chu C. 2020. The design and construction of reference pangenome graphs with
698 minigraph. *Genome Biology* **21**: 265.

699 Li J, Llorente B, Liti G, Yue J-X. 2022. RecombineX: A generalized computational framework
700 for automatic high-throughput gamete genotyping and tetrad-based recombination
701 analysis. *PLOS Genetics* **18**: e1010047.

702 Lian Q, Huettel B, Walkemeier B, Mayjonade B, Lopez-Roques C, Gil L, Roux F, Schneeberger
703 K, Mercier R. 2024. A pan-genome of 69 Arabidopsis thaliana accessions reveals a
704 conserved genome structure throughout the global species range. *Nat Genet* **56**:
705 982–991.

706 Liao W-W, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, Lu S, Lucas JK, Monlong J, Abel HJ,
707 et al. 2023. A draft human pangenome reference. *Nature* **617**: 312–324.

708 Liu H, Caballero-Florán RN, Hergenreder T, Yang T, Hull JM, Pan G, Li R, Veling MW, Isom LL,
709 Kwan KY, et al. 2023. DSCAM gene triplication causes excessive GABAergic synapses
710 in the neocortex in Down syndrome mouse models. *PLOS Biology* **21**: e3002078.

711 Miao Z, Ren Y, Tarabini A, Yang L, Li H, Ye C, Liti G, Fischer G, Li J, Yue J-X. 2024. ScRAPdb: an
712 integrated pan-omics database for the Saccharomyces cerevisiae reference assembly
713 panel. *Nucleic Acids Research* gkae955.

714 Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N,
715 Uralsky L, Gershman A, et al. 2022. The complete sequence of a human genome.
716 *Science* **376**: 44–53.

717 O'Donnell S, Yue J-X, Saada OA, Agier N, Caradec C, Cokelaer T, De Chiara M, Delmas S,
718 Dutreux F, Fournier T, et al. 2023. Telomere-to-telomere assemblies of 142 strains

719 characterize the genome structural landscape in Saccharomyces cerevisiae. *Nat*
720 *Genet* **55**: 1390–1399.

721 Paten B, Novak AM, Eizenga JM, Garrison E. 2017. Genome graphs and the evolution of
722 genome inference. *Genome Res* **27**: 665–676.

723 Raney BJ, Barber GP, Benet-Pagès A, Casper J, Clawson H, Cline MS, Diekhans M, Fischer C,
724 Navarro Gonzalez J, Hickey G, et al. 2024. The UCSC Genome Browser database: 2024
725 update. *Nucleic Acids Research* **52**: D1082–D1088.

726 Rech GE, Radío S, Guirao-Rico S, Aguilera L, Horvath V, Green L, Lindstadt H, Jamilloux V,
727 Quesneville H, González J. 2022. Population-scale long-read sequencing uncovers
728 transposable elements associated with gene expression variation and adaptive
729 signatures in Drosophila. *Nat Commun* **13**: 1948.

730 Salzberg LI, Martos AAR, Lombardi L, Jermiin LS, Blanco A, Byrne KP, Wolfe KH. 2022. A
731 widespread inversion polymorphism conserved among Saccharomyces species is
732 caused by recurrent homogenization of a sporulation gene family. *PLOS Genetics* **18**:
733 e1010525.

734 Sullivan MJ, Petty NK, Beatson SA. 2011. Easyfig: a genome comparison visualizer.
735 *Bioinformatics* **27**: 1009–1010.

736 Wagner J, Olson ND, Harris L, McDaniel J, Cheng H, Fungtammasan A, Hwang Y-C, Gupta R,
737 Wenger AM, Rowell WJ, et al. 2022. Curated variation benchmarks for challenging
738 medically relevant autosomal genes. *Nat Biotechnol* **40**: 672–680.

739 Weller CA, Andreev I, Chambers MJ, Park M, Program NCS, Bloom JS, Sadhu MJ, Barnabas
740 BB, Black S, Bouffard GG, et al. 2023. Highly complete long-read genomes reveal
741 pangenomic variation underlying yeast phenotypic diversity. *Genome Res* **33**: 729–
742 740.

743 Wick RR, Schultz MB, Zobel J, Holt KE. 2015. Bandage: interactive visualization of de novo
744 genome assemblies. *Bioinformatics* **31**: 3350–3352.

745 Yang X, Wang X, Zou Y, Zhang S, Xia M, Fu L, Vollger MR, Chen N-C, Taylor DJ, Harvey WT, et
746 al. 2023. Characterization of large-scale genomic differences in the first complete
747 human genome. *Genome Biology* **24**: 157.

748 Yokoyama TT, Sakamoto Y, Seki M, Suzuki Y, Kasahara M. 2019. MoMI-G: modular multi-scale
749 integrated genome graph browser. *BMC Bioinformatics* **20**: 548.

750 Yuan Y, Ma RK-K, Chan T-F. 2023. PanGraphViewer: A Versatile Tool to Visualize Pangenome
751 Graphs. 2023.03.30.534931.
752 https://www.biorxiv.org/content/10.1101/2023.03.30.534931v1 (Accessed July 10,
753 2024).

Yue J-X, Li J, Aigrain L, Hallin J, Persson K, Oliver K, Bergström A, Coupland P, Warringer J, Lagomarsino MC, et al. 2017. Contrasting evolutionary genome dynamics between domesticated and wild yeasts. *Nat Genet* **49**: 913–924.

Figure legends

Figure 1. The interactive user interface of VRPG when opened in a web browser. After selecting the input pangenome graph dataset, users can interactively visualize, navigate, and query the pangenome graph via a primary-linear-reference-based coordinate system. An example of the human *APOE* gene region (Chr19:44905000-44910000, GRCh38 coordinates) of the HRPC Minigraph-Cactus pangenome graph is shown here. The *APOE* bears strong associations with cardiovascular and Alzheimer's diseases. No layout simplification or node scaling was applied in VRPG for this visualization. Path highlighting for CHM13 was selected. All other options were left with defaults.

Figure 2. VRPG's visualization for different types of genomic variants. For different types of genomic variants such as (A) Divergent region, (B) Insertion, (C) Deletion, (D) Inversion, (E) Duplication, and (F) Duplication with divergent region, their VRPG-based graph representations are depicted with path highlighting feature enabled for the reference and query genome respectively. Information of the highlighted path regarding nodes' traversing order and orientation is further reported (in GAF-specified style) in VRPG's the path information panel.

Figure 3. VRPG's visualization for the Chromosome XIV (ChrXIV) flip/flop region in a yeast pangenome graph. (A) The genome sequence and synteny comparison among SGDref (S288C), Y12, and YPS128 for the ChrXIV flip/flop region, with the red and blue shades representing homologous regions shared with >97% sequence similarity (blue: same direction as shown in the flip/flop IR regions; red: reverse direction as shown in the flip/flop internal region). (B-D) The VRPG visualization for the graph in this region (ChrXIV:569857-602365, SGDref coordinates). The assembly paths of SGDref (B), Y12 (D) and YPS128 (E) are highlighted respectively, with the SGDref-based coordinate system and gene annotation track (C) further shown in between. No layout simplification or node scaling was applied in VRPG for this visualization. All other options were left with defaults.

Figure 4. VRPG's visualization for the polymorphic *DOG2* deletion in a yeast pangenome graph. (A) The genome sequence and synteny comparison among SGDref, W303, SK1 and Y12 for the *DOG2* region, with the blue shades representing homologous regions shared with >97% sequence similarity. (B-F) The VRPG visualization for the graph in this region (ChrVIII:189131-197233, SGDref coordinates). The assembly paths of S288C (B), W303 (D), Y12(E), and SK1 (F) are highlighted respectively, with the SGDref-based coordinate system and gene annotation track (C) further shown in between. The light purple shades denote the *DOG2* gene region. No layout simplification or node scaling was applied in VRPG for this visualization. All other options were left with defaults.

Figure 5. VRPG's visualization for the *DSCAM* intronic inversion in the HPRC human pangenome graphs. (A) The University of California, Santa Cruz (UCSC) Genome Browser view of the *DSCAM* exons 31-33 region at Chr21:40,010,001-40,045,000 (GRCh38 coordinates) with annotation tracks for chromosome ideogram, gene structure, repetitive sequence features displayed. (B) The genome sequence and synteny comparison between GRCh38 and CHM13 for the *DSCAM* exons 31-33 region, with the blue (matched in same directions) and red (matched in reversed directions) shades representing homologous regions shared with >98% sequence similarity. (C-E) VRPG visualization for the *DSCAM* inversion in the HPRC human pangenome graphs derived from Minigraph (C), Minigraph-Cactus (D), and PGGB (E), with the assembly paths of GRCh38 (left) and CHM13 (right) highlighted respectively. The light purple shades denote the *DSCAM* exons 31-33 region. The small

triangles along the graph indicate the trace of genomic variants shown in the graph, with those corresponding to the highlighted path further colored in red. VRPG's squeezed layout was used when visualizing the Minigraph-Cactus and PGGB graphs. Layout simplification was applied as indicated, whereas node scaling was enabled in all cases. All other options were left with defaults.

Figure 6. VRPG's visualization for the *CR1* intragenic deletion in the HPRC human pangenome graphs. (A) The UCSC Genome Browser view for the *CR2*-*CR1*-*CR1L* region at Chr1:207,453,024-207,738,416 (GRCh38 coordinates) with annotation tracks for chromosome ideogram, gene structure, repetitive sequence features displayed. (B) The genome sequence and synteny comparison between GRCh38 and CHM13 for the *CR2*-*CR1*-*CR1L* region, with the blue shades representing homologous regions shared with >98% sequence similarity. The deletion is further highlighted in the red triangle. (C-E) VRPG's visualization for the *CR1* deletion in the HPRC human pangenome graphs derived from Minigraph (C), Minigraph-Cactus (D), and PGGB (E), with the genome paths of GRCh38 and CHM13 highlighted respectively. The pink shades denote the *CR1* genic region. The path differences in the deletion region are further indicated by black triangles. VRPG's squeezed layout was used for the Minigraph-Cactus and PGGB graph visualization. Nonref node simplification was applied as indicated when visualizing the Minigraph-Cactus and PGGB graphs. All other options were left with defaults.

Tables

Table 1. Basic information and functionality comparison between VRPG and other pangenome graph visualization tools.

	VRPG	**Bandage**	**GfaViz**	**gfaestus**	**Sequence TubeMap**	**MoMI-G**	**ODGI**[1]	**PGR-TK**[1]	**PanGraphViewer**
Distributing license	MIT	GNU GPL v3	ISC	MIT	MIT	MIT	MIT	MIT	MIT
Supported operating system	Linux	Linux; MacOS; Windows	Linux; MacOS; Windows	Linux; MacOS	Linux; MacOS	Linux; MacOS	Linux	Linux; MacOS	Linux; MacOS; Windows
Supported pangenome graph format	GFAv1; rGFA	GFAv1; rGFA;	GFAv1; GFAv2; rGFA	GFAv1 (also needs ODGI's layout TSV)	xg; vg; gbz	xg	GFAv1; og	GFAv1	GFAv1; rGFA
Deployment	Web-based	Desktop-based	Desktop-based	Desktop-based	Web-based	Web-based	Desktop-based	Desktop-based	Web-based; Desktop-based
Rendering object	Subgraph	Full graph	Full graph	Full graph	Subgraph	Subgraph	Subgraph	Subgraph	Subgraph
Graph query mode	By coordinates	By node IDs	By node IDs	By node IDs	By coordinates; By node IDs	By coordinates	By coordinates	N/A	By coordinates
Rendering style	Dynamic	Dynamic	Dynamic	Dynamic	Dynamic	Dynamic	Static	Static	Dynamic
Support for annotation[2]	BED; GFF3	TSV	N/A	BED; TSV	BED	BED; GFF3; WIG	N/A	N/A	BED; GFF3; GTF
Sequence-to-graph mapping	Yes (for graphs built by Minigraph)	Yes	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Assembly-to-graph path highlighting	Yes	N/A	N/A	N/A	N/A	N/A	Yes (partial)	N/A	N/A
Assembly-specific node depth report	Yes	N/A	N/A	N/A	N/A	N/A	Yes (bin-based)	N/A	N/A

Footnote:

1: Tools like ODGI and PGR-TK are multi-purpose software suites, both of which perform graph visualization via their submodules. The function comparison list here only concerns their visualization module.

2: Although multiple tools accept user-supplied annotation files, it varies considerably regarding how they utilize these files with the graph.

Table 2. Performance comparison between VRPG and other pangenome graph visualization tools.

Benchmarking task	Graph source	Graph type	Graph builder	Graph complexity (N:node; E:edge)	VRPG[1]	Bandage	GfaViz	PanGraphViewer
Graph parsing	Yeast	Full-graph	Minigraph	N: 37,062 E: 52,756	0.10 sec 64.80 MB	3.50 sec 254.11 MB	247.11 sec 2,769.50 Mb	0.22 sec 282.10 MB
Graph parsing	Human	Full-graph	Minigraph	N:424,643 E: 637,628	0.07 sec 54.44 MB	80.35 sec 10,066.50 MB	NA (memory cap)	7.83 sec 524.38 MB
			Minigraph-Cactus	N: 81,415,956 E: 112,955,105	0.10 sec 64.73 MB	NA (memory cap)	NA (memory cap)	NA (parsing error)
			PGGB	N: 110,884,673 E: 154,756,169	0.09 sec 58.26 MB	NA (memory cap)	NA (memory cap)	NA (parsing error)
Graph parsing	Human	Subgraph> Chr22	Minigraph	N: 9,545 E:14,307	0.01 sec 62.5MB	1.80 sec 260.15 MB	242.35 sec; 10,083.49 MB	0.19 sec 307.11 MB
			Minigraph-Cactus	N: 1,279,308 E: 1,780,703	0.01 sec 58.16 MB	84.50 sec 2,120.66 MB	NA (memory cap)	7.88 sec 815.58 MB
			PGGB	N: 1,124,300 E: 1,568,024	0.02 sec 56.11 MB	71.05 sec 1,890.45 MB	NA (memory cap)	NA (parsing error)
Graph rendering	Yeast	Subgraph> ChrIV:1-50000	Minigraph	N:103 E:142	0.02 sec 82.96 MB	0.76 sec 105.94 MB	0.12 sec 171.48 MB	1.08 sec 323.73 MB
		Subgraph> ChrIV:1-500000	Minigraph	N:1,155 E:1,641	0.02 sec 194.86 MB	1.51 sec 122.36 MB	3.66 sec 324.79 MB	1.68 sec 324.70 MB
Graph rendering	Human	Subgraph> Chr22:2000001-2100000	Minigraph	N:13 E:18	0.01 sec 55.24 MB	0.30 sec 103.70 MB	0.16 sec 171.23 MB	0.01 sec 255.26 MB
			Minigraph-Cactus	N: 2,248 E: 3,106	0.05 sec 219.78 MB	1.81 sec 129.66 MB	13.25 sec 772.58 MB	1.88 sec 339.91 MB
			PGGB	N: 2,252 E: 3,116	0.05 sec 237.48 MB	1.68 sec 129.81 MB	14.15 sec 785.69 MB	NA (parsing error)
		Subgraph> Chr22:20000001-20500000	Minigraph	N:288 E:435	0.03 sec 82.78 MB	1.33 sec 111.81 MB	0.36 sec 181.08 MB	1.29 sec 321.28 MB
			Minigraph-Cactus	N: 21,982 E: 30,639	0.18 sec 605.82 MB	19.12 sec 347.27 MB	NA (memory cap)	7.37 sec 338.50 MB
			PGGB	N: 66,132 E: 94,595	0.33 sec 1,270.61 MB	69.20 sec 766.10 MB	NA (memory cap)	NA (parsing error)

Footnote:

1: The format conversion (for Minigraph-Cactus and PGGB graphs) and index building step were performed in advance and not included in this benchmarking.

PangenomeGraph Information Help

An interactive visualization and interpretation framework for linear-reference-projected pangenome graphs

① Select the pangenome graph for visualization

Pangenome graph dataset: HPRC Minigraph-Cactus graph (90 Homo sapiens assemblies)

Specify the primary linear-reference-based query region

chr19 | 44905000 - 44910000 | Search depth: 10 | Go

Layout: Ultra expanded | Simplify: None | Min bubble size: 50 | Node: Nonscaled | Edge: Straight

Path highlighting: 1 selected | Sequence-to-graph mapping | Stop optimization | Save image (*.svg)

② Select preferred graph simplification strategy & parameter

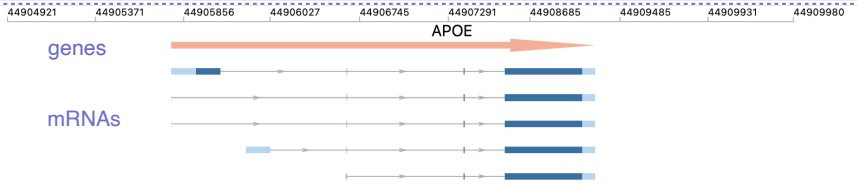
Select assembly for path highlighting

Pangenome graph

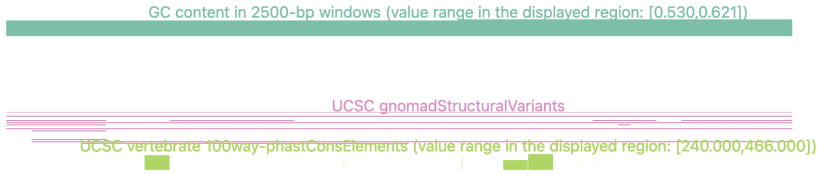
Mouse-selected query node

Highlighted path for the selected assembly: CHM13#0

③ Gene annotation



④ Other annotations



⑤ Information of the selected node

Node information

[Download]

Node name:
32987709

Node source:
Assembly: CHM13#0
Chr: chr19
Start: 47730606
End: 47730606
Length: 1

Node sequence:

Gene list:

ID	Name
gene:ENSG05220021152	*APOE*

Node depth:

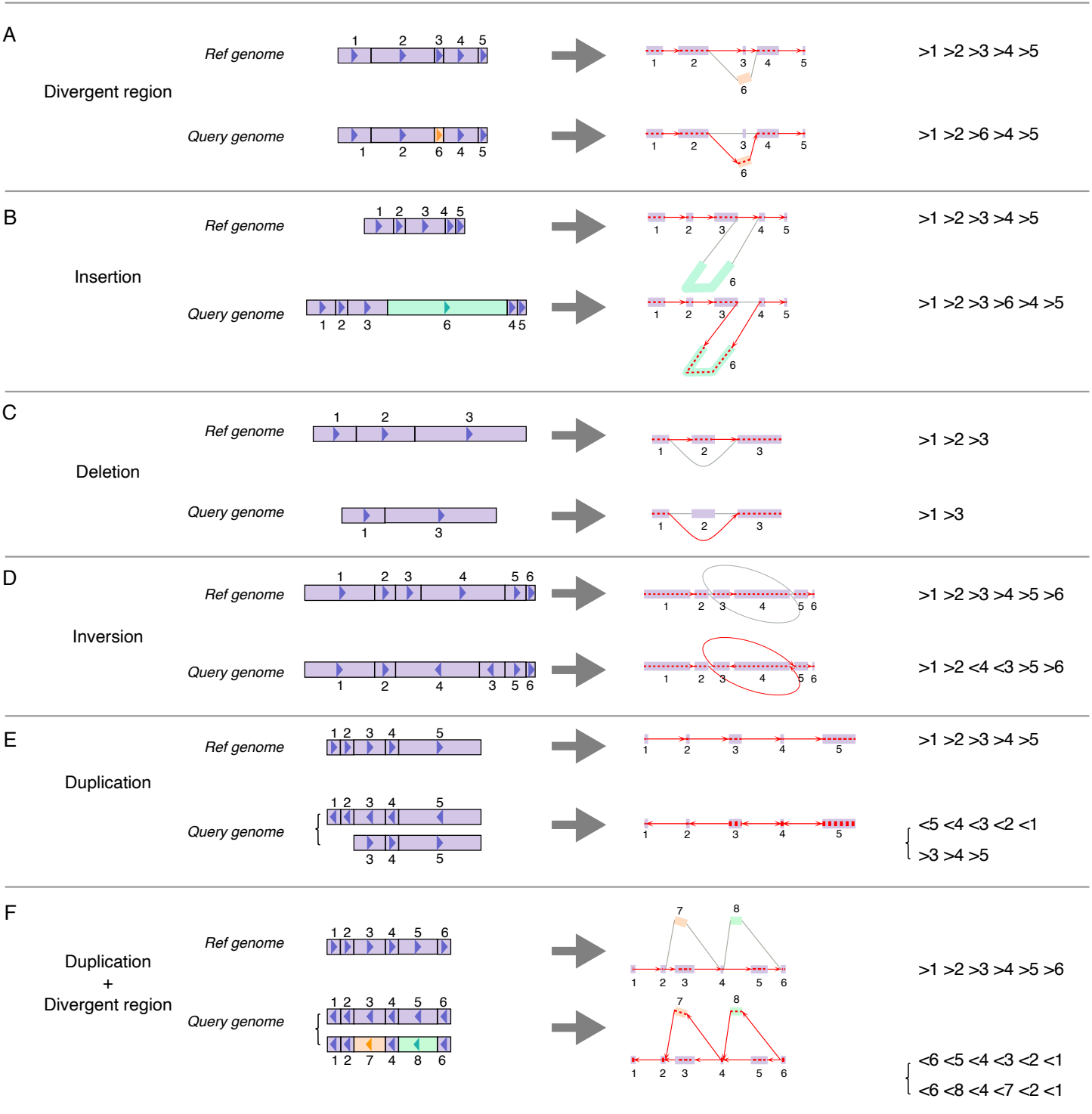
Assembly	Depth
GRCh38#0	0
CHM13#0	1
HG00438#1	1

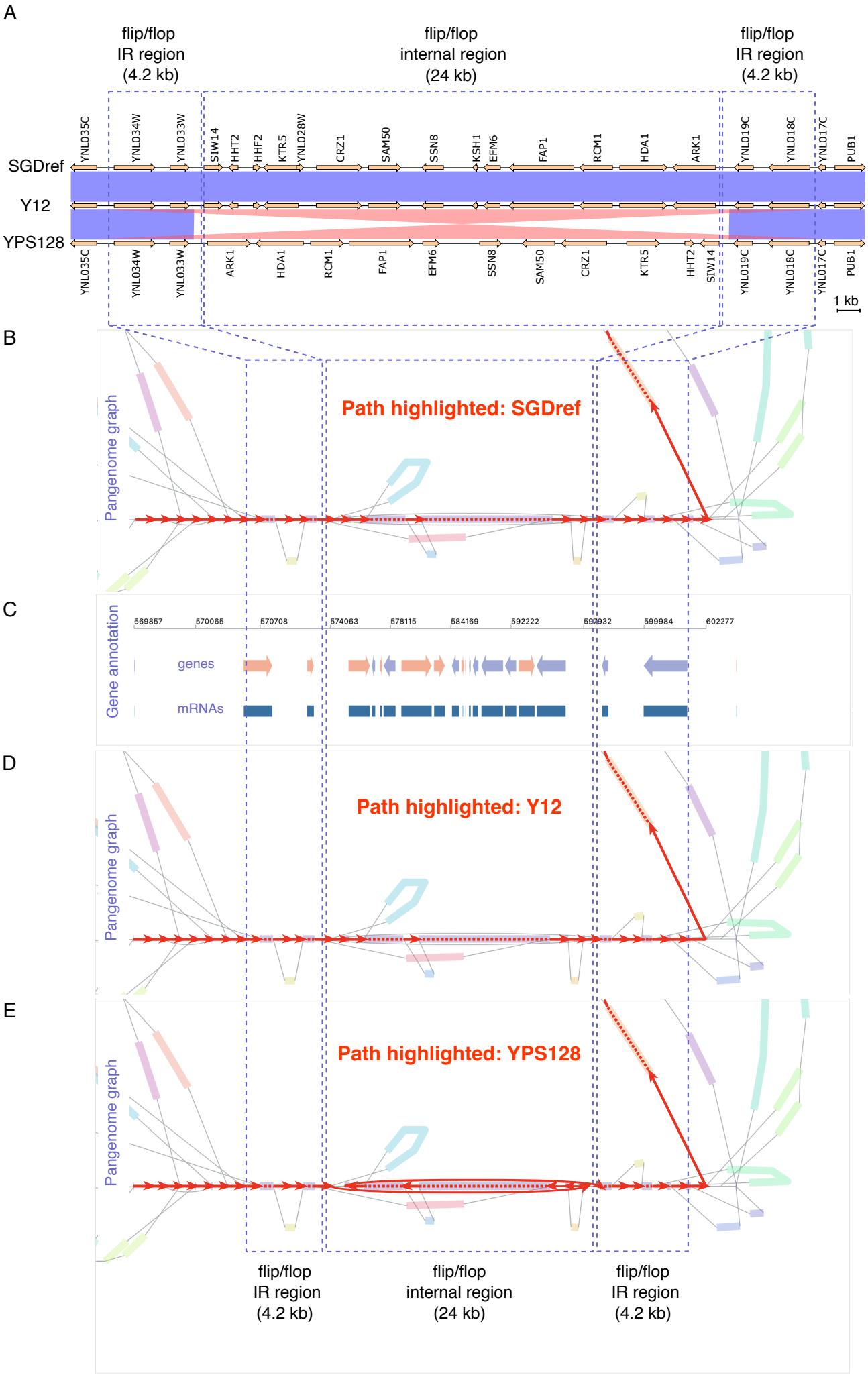
Scrollable

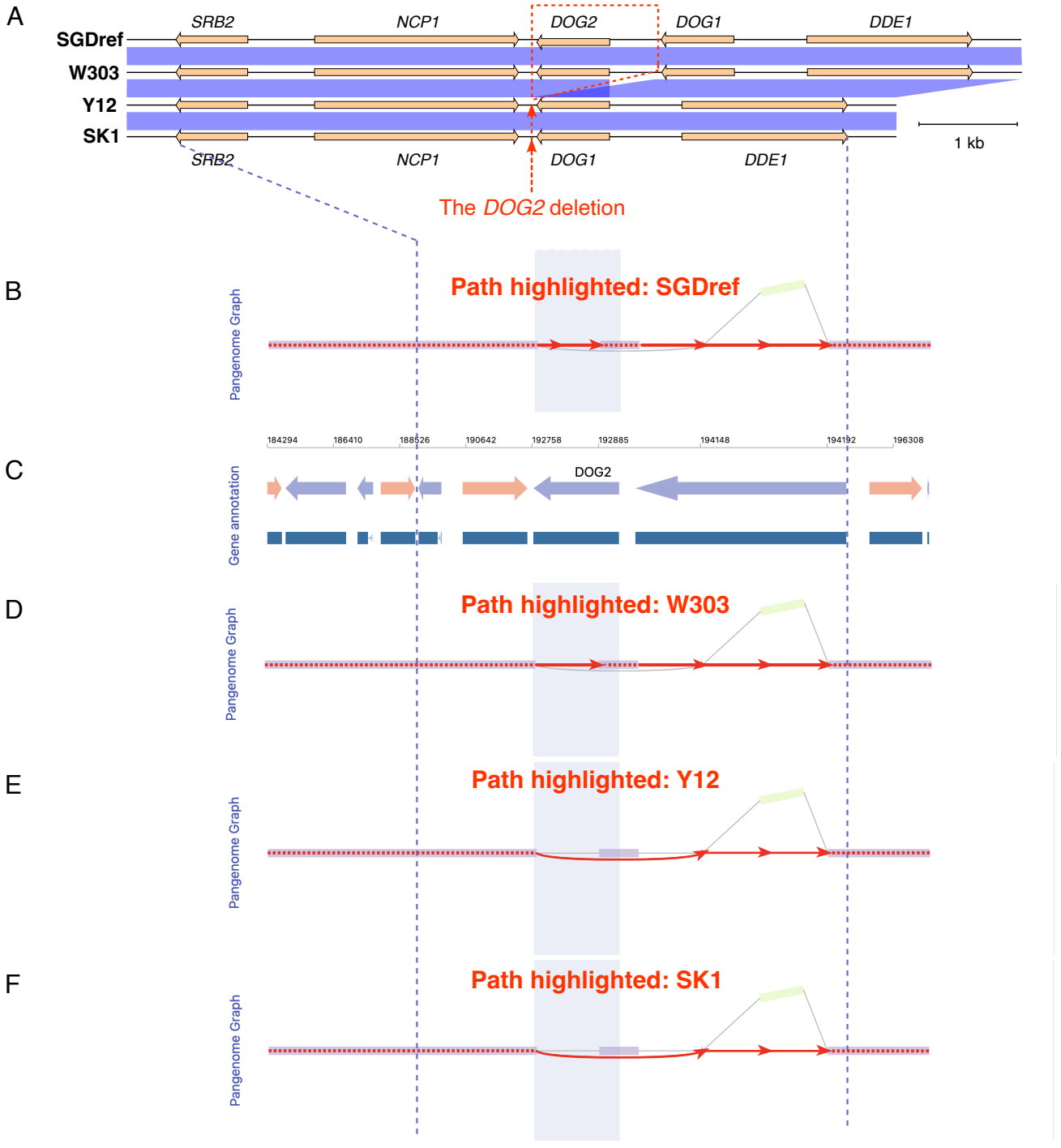
⑥

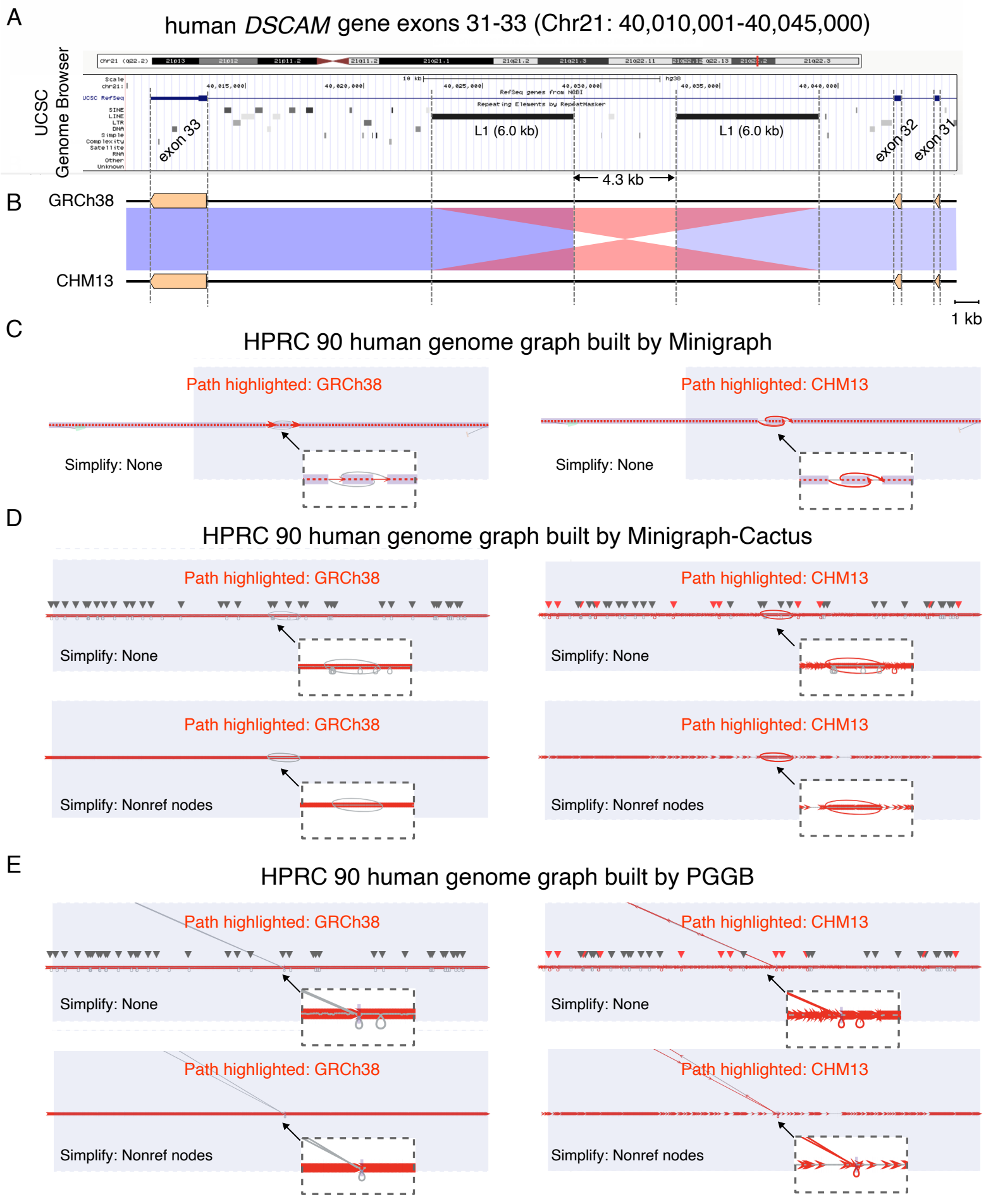
qChr	qStart	qEnd	qPath
CHM13#0#chr19	47729617	47734675	>32987684>32987685>32987687>32987688>32987690>32987691>32987693>32987694>32987696>32987698>32987699

Genome coordinates & in-graph path for the highlighted assembly


Variant type
Linear representation
Graph representation in VRPG
Highlighted path
A
Divergent region
Ref genome
Query genome
>1 >2 >3 >4 >5
>1 >2 >6 >4 >5
B
Insertion
Ref genome
Query genome
>1 >2 >3 >4 >5
>1 >2 >3 >6 >4 >5
C
Deletion
Ref genome
Query genome
>1 >2 >3
>1 >3
D
Inversion
Ref genome
Query genome
>1 >2 >3 >4 >5 >6
>1 >2 <4 <3 >5 >6
E
Duplication
Ref genome
Query genome
>1 >2 >3 >4 >5
<5 <4 <3 <2 <1
>3 >4 >5
F
Duplication
+
Divergent region
Ref genome
Query genome
>1 >2 >3 >4 >5 >6
<6 <5 <4 <3 <2 <1
<6 <8 <4 <7 <2 <1

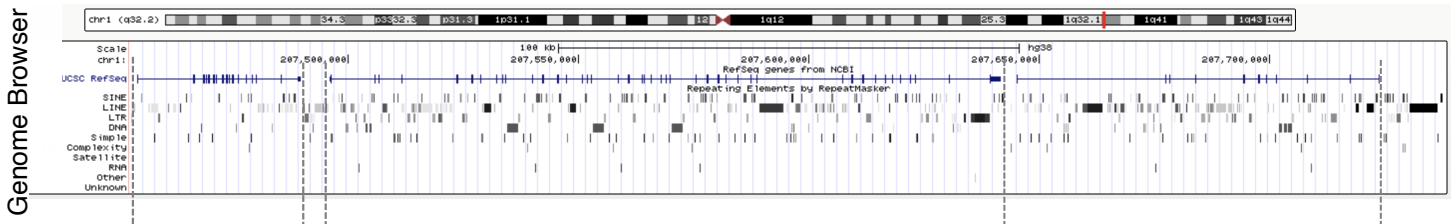

A
flip/flop
IR region
(4.2 kb)
flip/flop
internal region
(24 kb)
flip/flop
IR region
(4.2 kb)
SGDref
Y12
YPS128
YNL035C
YNL034W
YNL033W
SIW14
HHT2
HHF2
KTR5
YNL028W
CRZ1
SAM50
SSN8
KSH1
EFM6
FAP1
RCM1
HDA1
ARK1
YNL019C
YNL018C
YNL017C
PUB1
1 kb
B
Pangenome graph
Path highlighted: SGDref
C
Gene annotation
569857
570065
570708
574063
578115
584169
592222
597932
599984
602277
genes
mRNAs
D
Pangenome graph
Path highlighted: Y12
E
Pangenome graph
Path highlighted: YPS128
flip/flop
IR region
(4.2 kb)
flip/flop
internal region
(24 kb)
flip/flop
IR region
(4.2 kb)


A
SRB2
NCP1
DOG2
DOG1
DDE1
SGDref
W303
Y12
SK1
1 kb
The DOG2 deletion
B
Path highlighted: SGDref
Pangenome Graph
184294
186410
188526
190642
192758
192885
194148
194192
196308
C
Gene annotation
DOG2
D
Path highlighted: W303
Pangenome Graph
E
Path highlighted: Y12
Pangenome Graph
F
Path highlighted: SK1
Pangenome Graph

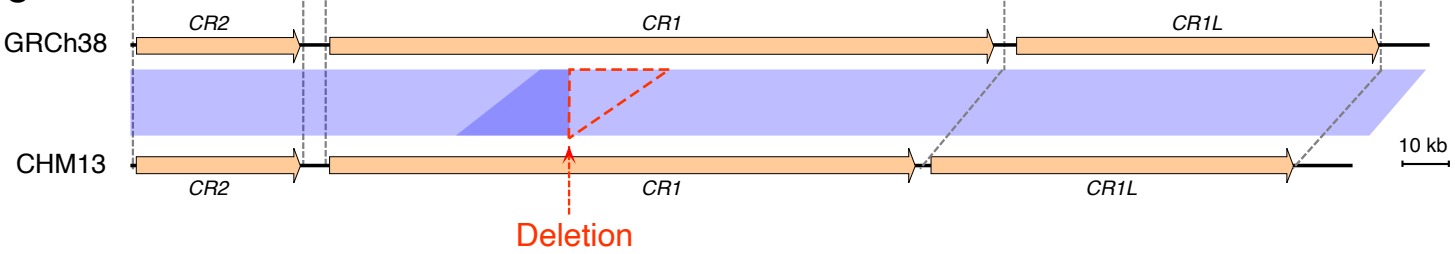

A
human DSCAM gene exons 31-33 (Chr21: 40,010,001-40,045,000)
UCSC Genome Browser
exon 33
L1 (6.0 kb)
L1 (6.0 kb)
exon 32
exon 31
4.3 kb
B
GRCh38
CHM13
1 kb
C
HPRC 90 human genome graph built by Minigraph
Path highlighted: GRCh38
Simplify: None
Path highlighted: CHM13
Simplify: None
D
HPRC 90 human genome graph built by Minigraph-Cactus
Path highlighted: GRCh38
Simplify: None
Path highlighted: CHM13
Simplify: None
Path highlighted: GRCh38
Simplify: Nonref nodes
Path highlighted: CHM13
Simplify: Nonref nodes
E
HPRC 90 human genome graph built by PGGB
Path highlighted: GRCh38
Simplify: None
Path highlighted: CHM13
Simplify: None
Path highlighted: GRCh38
Simplify: Nonref nodes
Path highlighted: CHM13
Simplify: Nonref nodes

A

human *CR2-CR1-CR1L* region (Chr1: 207,453,024-207,738,416)

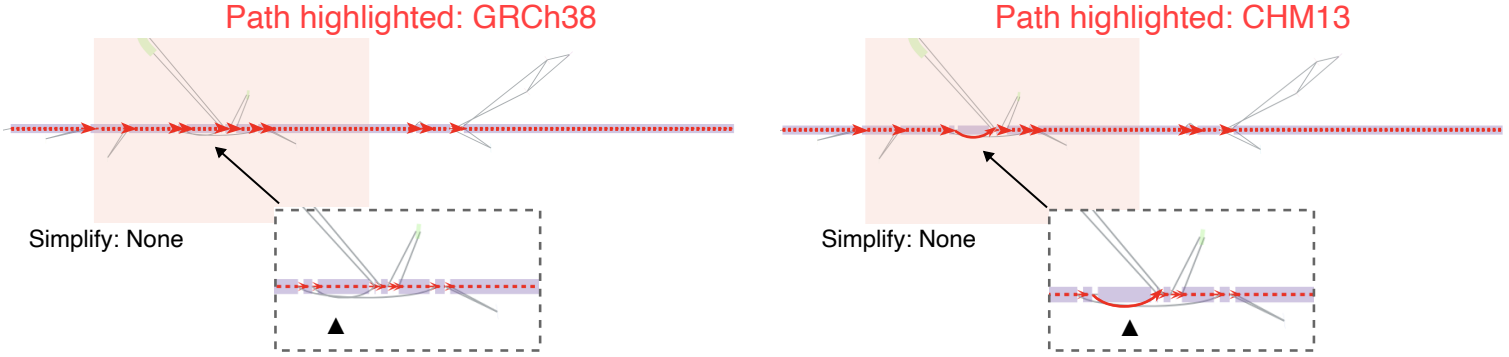


B



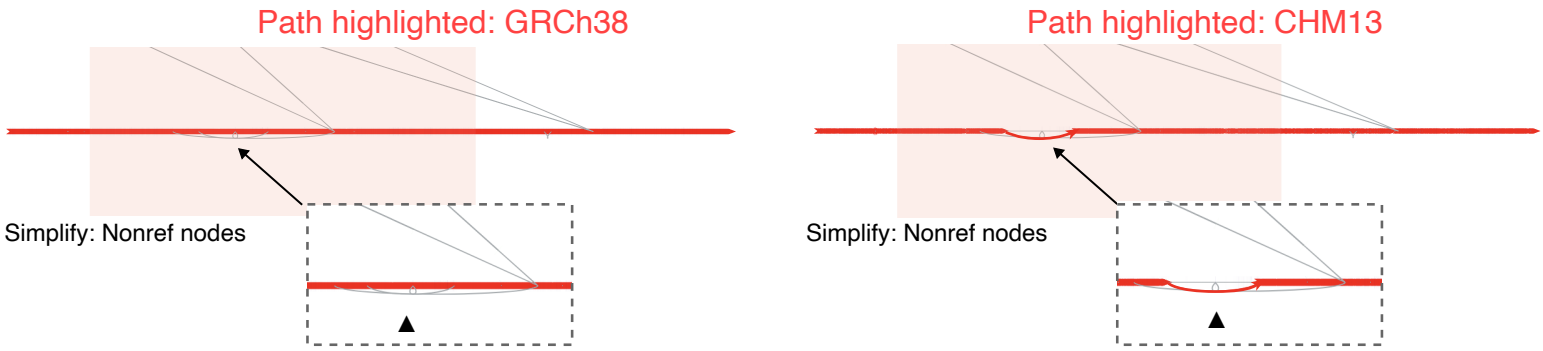
C

HPRC 90 human genome graph built by Minigraph



D

HPRC 90 human genome graph built by Minigraph-Cactus



E

HPRC 90 human genome graph built by PGGB