Variation graph applications

Dissertation
der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Sebastian Vorbrugg
aus Baden-Baden

Tübingen
2025

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	28.10.2025
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter/-in:	Prof. Dr. Detlef Weigel
2. Berichterstatter/-in:	Prof. Dr. Daniel Huson

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

Sebastian Vorbrugg
February 2026

Acknowledgements

First, I want to thank Prof. Detlef Weigel. He provided me with excellent supervision, support, and freedom throughout my Ph.D. I truly enjoyed being part of your team!

I also want to thank my office mates during this time: Christian, Max, Wenfei, Zhigui, Ilja, Oliver, Pablo, Shanshan, Wei, and Lana. We always had a great time together, with interesting discussions, relaxed conversations on the side, and a pleasant working atmosphere.

A special thanks to Christian and Max, who took me under their wings when I first joined the lab. I'm also grateful to everyone who participated in the monthly genome informatics meetings, especially Haim and Fernando, for their valuable feedback on my projects over the years. I particularly appreciate both of them for enduring the challenges of the 1001G+ projects with me, sharing insights, staying committed, and pushing through to the end. I'm thankful it's finally done!

I want to thank Rebecca for taking me on one of the sampling trips to Sweden during a difficult time in my work. It was a fun and eye-opening experience where I learned a lot and saw firsthand how the data I was working with had been collected in the years before. It was such an enjoyable and inspiring work trip that I could imagine doing that kind of fieldwork all the time.

Very special thanks to Ilja, without you, I would never have finished my Ph.D. Thank you for your constant motivation on my first paper and for showing me that sometimes just submitting can be harder than doing the research. I appreciate all your support, great advice, and honest feedback. I also enjoyed our chats about bike rides, cycling gear, and casual sports. Thank you!

I also want to thank my girlfriend, Lioba, for her heartfelt support throughout my entire Ph.D., for sharing a home office with me, for all the trips and vacations we took during that time, and for being such a wonderful partner. Thank you for commuting two hours every day just to live with me in the Tübingen.

Finally, I want to thank my parents for being so supportive and for giving me the freedom to study whatever I wanted, without any financial pressure. I couldn't have done this without you.

Abstract

Variation graphs provide a powerful solution to overcome the limitations of linear reference genomes, especially in representing the diversity and structural complexity within species. As genome sequencing becomes more accessible and datasets grow in both quality and scope, it is increasingly clear that traditional reference-based analyses fall short in capturing large-scale variation, population structure, and genomic complexity. However, the practical interpretation and use of genome graphs remains an open challenge. Both graph construction and downstream analysis require new tools that can operate at scale, preserve biological interpretability, and offer meaningful metrics to describe the underlying structure.

In this thesis, I present a set of tools developed to address key challenges in variation graph analysis. The core contribution is `gretl`, a fast and flexible framework for computing graph- and path-based statistics. It enables systematic comparisons across parameter settings and graph construction methods, and has been used to analyze graphs built from multiple species, including a yeast dataset and the 1001 Genomes *Arabidopsis* pangenome. The framework reveals how parameters such as segment length and alignment thresholds strongly affect graph structure and interpretability. I also introduce `gfa2bin`, a graph-to-GWAS bridge that supports association testing directly from graph node coverage. This method demonstrates the potential of graph-based GWAS to detect both known and novel signals of trait associations. In addition, I develop a novel variation detection approach based on bifurcation events between paths, offering a complementary alternative to standard bubble detection algorithms.

Together, these tools enable direct statistical exploration and biological analysis of genome graphs at both global and sample-specific levels. Applied to the *Arabidopsis* dataset, they reveal population structure, patterns of pangenome expansion, and the role of private and structural variation across diverse accessions. While challenges remain in variant extraction, graph augmentation, and performance scaling, this work demonstrates that genome graphs can be used not only to store variation, but also to interpret and analyze it in meaningful ways. The tools and methods presented here are a step toward more flexible, interpretable, and biologically aware graph-based genomics.

Zusammenfassung

Die Analyse genetischer Variation ist ein zentrales Thema der modernen Genomforschung. Klassische, lineare Referenzgenome stoßen dabei zunehmend an ihre Grenzen, insbesondere, wenn es um die Darstellung struktureller Vielfalt innerhalb einer Art geht. Variationsgraphen bieten hier einen innovativen Ansatz: Sie ermöglichen eine umfassendere Repräsentation genetischer Diversität und erlauben die direkte Modellierung komplexer struktureller Varianten.

Mit dem Zugang zu hochwertigen Sequenzierungsdaten gewinnt dieses zunehmend Modell an Bedeutung. Gleichzeitig fehlen jedoch oft geeignete Tools, um solche Graphstrukturen systematisch zu analysieren und biologisch zu interpretieren.

In dieser Arbeit wurden mehrere Methoden entwickelt, die zentrale Herausforderungen bei der Analyse von Variationsgraphen adressieren. Das Tool `gretl` bietet eine flexible Plattform zur Berechnung verschiedener struktur- und pfadbasierter Metriken und erlaubt den Vergleich unterschiedlicher Parameter und Konstruktionsmethoden. Es wurde unter anderem auf einem *Saccharomyces cerevisiae*-Datenset sowie auf dem 1001-Genomes-Pangenom von *Arabidopsis thaliana* angewendet. Dabei zeigte sich, dass Parameter wie Segmentlänge die Topologie und Interpretierbarkeit der resultierenden Graphen maßgeblich beeinflussen.

Mit `gfa2bin` wurde darüber hinaus ein Tool für graphbasierte Assoziationsanalysen (GWAS) entwickelt. Es verwendet knotenbasierte Coverage-Werte als Grundlage für die Genotypisierung, die anschließend für statistische Tests herangezogen werden. Dabei konnten sowohl bekannte als auch potenziell neue Assoziationen identifiziert werden. Ergänzt wird das Methodenset durch einen neuartigen Ansatz zur Variantendetektion, der auf Verzweigungen zwischen Pfaden basiert und klassische Bubble-Erkennungsverfahren gezielt erweitert.

Die vorgestellten Methoden ermöglichen neue Einblicke in die Struktur, Variation und funktionellen Zusammenhänge von Genomgraphen. Am Beispiel von *Arabidopsis thaliana* lassen sich sowohl Populationsstrukturen als auch das Wachstum des Pangenoms detailliert nachvollziehen. Trotz bestehender Herausforderungen, etwa bei der effizienten Variantenerkennung, der Integration neuer Daten in bestehende Graphen oder dem Umgang mit großen Datensätzen, zeigen die Ergebnisse deutlich: Variationsgraphen haben das Potenzial, nicht

nur als Repräsentation, sondern als aktives Werkzeug zur Analyse genetischer Variation etabliert zu werden.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

Deoxyribonucleic acid (DNA) is the fundamental hereditary material in all known living organisms. It serves as the primary repository of genetic information, encoding the instructions necessary for growth, development, function, and reproduction of cells. The specific sequence of nucleotide bases, adenine (A), thymine (T), cytosine (C), and guanine (G), constitutes the genetic code, which directs the synthesis of proteins and regulates the expression of genes. As such, DNA plays a central role in maintaining the integrity and continuity of life across generations.

A foundational framework for understanding the flow of genetic information is articulated by the Central Dogma of Molecular Biology, a concept first proposed by Francis Crick in 1958 [22]. This principle describes the flow of genetic information as proceeding from DNA to RNA to protein, and fundamentally excludes the possibility of information transfer from protein back to nucleic acids. The process begins with transcription, in which a specific DNA sequence is transcribed into messenger RNA (mRNA). This mRNA then serves as a template for translation, during which ribosomes synthesize proteins by decoding the mRNA sequence into a corresponding chain of amino acids.

The vast majority of enzymatic reactions, including many steps of signal transduction as well as biosynthesis of metabolites is carried out by proteins. The accurate transmission of information from DNA to protein is therefore essential for normal cellular function and organismal development.

Thus, the Central Dogma provides not only a molecular framework for gene expression but also a mechanistic link between genotype and phenotype, an essential connection for interpreting evolutionary processes at the population level. It forms the bridge between molecular biology and evolutionary theory, enabling an integrated understanding of how genetic information translates into observable traits subject to evolutionary change.

In this context, the ability to sequence DNA across individuals and populations has be-

come a foundation of modern biological research. The development of high-throughput and cost-efficient sequencing technologies has transformed what was once a laborious and expensive task into a routine and scalable approach for exploring genetic diversity, evolutionary dynamics, and the molecular basis of complex traits.

1.1 DNA

The molecular story of DNA began in 1869, when Swiss biochemist Friedrich Miescher, working at the University of Tübingen, first isolated a substance from the nuclei of white blood cells, which he termed "nuclein" [93]. Although Miescher identified this material as chemically distinct, its role in heredity remained unknown. For decades, proteins were believed to be the likely carriers of genetic information due to their structural complexity. This assumption was challenged in 1944, when Oswald Avery, along with Colin MacLeod and Maclyn McCarty, demonstrated that DNA, not protein, was the "transforming principle" responsible for heredity in bacteria [8]. Their work provided the first strong evidence that DNA carried genetic information. A decisive breakthrough came in 1953, when James Watson and Francis Crick, drawing on crucial X-ray crystallography data from Rosalind Franklin, proposed the double-helix structure of DNA [139]. This model elegantly explained how DNA could be replicated and how genetic information is encoded in base sequences.

1.2 History of DNA sequencing

The history of DNA sequencing began with slow and technically demanding methods that relied on radioactive labeling and manual gel electrophoresis. Early efforts produced only short sequences, often requiring years of work to resolve fragments of a few dozen base pairs. A major breakthrough came in the late 1970s with the introduction of the Sanger sequencing method, which used chain-terminating nucleotides [120]. This approach, later automated and refined with fluorescent detection, enabled the first large-scale sequencing projects, including the Human Genome Project, though at great cost and effort [72, 136]. The development of next-generation sequencing (NGS) in the early 2000s marked a turning point [92]. By massively parallelizing sequencing reactions and reducing reagent volumes, NGS technologies dramatically increased throughput and lowered costs. This made whole-genome sequencing accessible and routine across many fields.

More recently, third-generation sequencing technologies have enabled real-time sequencing of single DNA molecules, producing much longer reads and offering new insights into structural variation and complex regions of the genome [31, 95]. While these methods can

have higher error rates, they are continually improving and are increasingly supported by complementary computational and experimental techniques [123].

1.3 Early approaches

Early methods focused on RNA viruses [34]. The first complete protein-coding gene gene sequences was published by Walter Fiers and colleagues, who used enzymatic digestion and two-dimensional gel electrophoresis to determine the RNA sequence of the bacteriophage MS2 [34]. These methods were technically complex, slow, and limited to short fragments, but demonstrated that nucleotide sequences could be resolved experimentally with biochemical tools.

A more scalable approach was introduced by Ray Wu, who developed a technique based on DNA polymerase extension from a known primer, incorporating radiolabeled nucleotides [144]. This marked a conceptual shift toward synthesis-based sequencing and laid the groundwork for the first broadly adopted DNA sequencing method.

In 1977, Frederick Sanger published the chain-termination method, which became the foundation of first-generation sequencing. This technique used dideoxynucleotides (ddNTPs) to terminate DNA synthesis at specific bases [120]. By performing four parallel reactions, one for each nucleotide, and separating the resulting fragments by polyacrylamide gel electrophoresis, it became possible to infer DNA sequences from the banding pattern. Later adaptations using fluorescently labeled terminators enabled automated detection and formed the basis of sequencing systems developed in the 1980s and 1990s [131].

Sanger sequencing proved highly accurate, with typical read lengths of 500–1000 base pairs, and remained the dominant technique in molecular biology for nearly three decades. However, it required clonal input DNA, typically generated through bacterial transformation. To sequence larger genomic regions, methods such as bacterial artificial chromosomes (BACs) and yeast artificial chromosomes (YACs) were developed to propagate long inserts [13, 106, 128].

Sanger sequencing played a central role in early large-scale genome projects, including the Human Genome Project [72, 136]. However, its low throughput and high cost made it unsuitable for sequencing many genomes or analyzing large populations efficiently. Its reliance on gel-based separation and single-fragment workflows limited scalability and speed, particularly as researchers sought to sequence entire genomes across a diverse range of species. High-profile projects such as the *Arabidopsis thaliana* genome [6], the first fully sequenced plant genome, and the *Drosophila melanogaster* genome [5], a model for animal genetics, also relied on Sanger-based approaches. Similar strategies were used to complete

the rice genome (*Oryza sativa*) [44, 145], a major milestone in crop genomics. These and other efforts made clear the need for faster, more scalable sequencing technologies, ultimately motivating the development of high-throughput sequencing platforms in the early 2000s.

Nevertheless, Sanger sequencing remains widely used today for targeted sequencing and validation, valued for its reliability and interpretability [87]. It also serves as a benchmark for assessing the accuracy of newer sequencing methods [126].

1.4 Next-Generation Sequencing

The limitations of first generation sequencing, particularly its low throughput and high cost, motivated the development of fundamentally new approaches to DNA sequencing [87, 92]. Around the turn of the 21st century, a range of technologies emerged that collectively came to be known as next generation sequencing (NGS). These methods introduced several important innovations: the ability to perform massively parallel sequencing reactions, the use of clonally amplified DNA fragments immobilised on solid supports, and real-time monitoring of sequencing-by-synthesis or related chemistries [127]. The result was a significant reduction in sequencing costs, a dramatic increase in output, and the establishment of a new standard in genomic research [126].

A key conceptual advance came from the work of George Church and colleagues, who demonstrated that DNA sequences could be read directly from spatially resolved DNA colonies, referred to as polonies, embedded in a polyacrylamide gel [97]. In parallel, advances in fluorescence microscopy, combined with increasingly high-resolution sensors, enabled the detection of base incorporation events. This early demonstration of parallelised sequencing-by-synthesis established the foundational principles for later developments in the field. Around the same time, Balasubramanian and Klenerman, through Solexa, a company later acquired by Illumina, introduced a more scalable sequencing implementation based on bridge amplification and reversible terminator chemistry [118, 11]. In this system, short DNA fragments ligated to surface-bound adapters are clonally amplified to form dense clusters on a flow cell. During sequencing, a DNA polymerase incorporates fluorescently labelled and chemically blocked nucleotides one at a time. After imaging, the blocking groups are removed, and the process repeats. This approach allowed millions of DNA fragments to be sequenced simultaneously with high accuracy and has since become the dominant platform for short-read sequencing [48].

Several other sequencing systems were developed in parallel. The pyrosequencing approach commercialised by 454 Life Sciences detected pyrophosphate release during nucleotide incorporation using a luciferase-based light detection system [88]. It was among the

first methods to produce longer sequencing reads but suffered from errors in homopolymeric regions and high costs. Ion semiconductor sequencing, developed by Ion Torrent, used pH sensors to detect hydrogen ions released during nucleotide incorporation. Although this method reduced reagent cost and increased speed, its accuracy and consistency did not match that of Illumina-based systems [118]. Helicos Biosciences pursued single-molecule sequencing using fluorescence detection, which avoided amplification entirely but faced technical challenges and limited commercial success [48].

The success of Illumina's platform is attributed to its combination of low per-base error rates, a relatively uniform error profile, and compatibility with a wide range of applications [11]. Despite these advantages, the technology is constrained by short read lengths, typically between 100 and 250 base pairs. This limits its ability to resolve structural variants, repetitive regions, and long-range haplotypes. In response, several experimental and computational strategies have been developed to extend the utility of short-read data. These include mate-pair libraries and synthetic long-read technologies such as Moleculo and 10X Genomics, which associate barcodes with long DNA fragments prior to fragmentation [147]. This enables the reconstruction of long-range information from short sequencing reads. Other strategies include strand-specific techniques such as strand-seq, which isolates DNA from individual chromatids [33], and proximity ligation methods such as Hi-C, which provide structural information by capturing spatial interactions between DNA regions [82]. These methods have supported improved haplotype resolution, structural variant detection, and genome assembly, especially in complex or previously uncharacterised genomes.

Next generation sequencing has had a profound impact on biological research. It enabled large-scale studies in genomics, transcriptomics, and metagenomics, and made high-throughput DNA sequencing accessible to a broad scientific community. Although subsequent technologies have addressed some of the limitations of short-read sequencing directly, next generation sequencing remains a widely used and continually evolving platform due to its scalability, accuracy, and adaptability.

1.5 "Third generation" sequencing

While second generation sequencing brought high-throughput and cost-effective DNA analysis, it is limited by its reliance on short reads and the observation of pooled molecules. Amplification, which is used to enhance signal, introduces potential biases and errors, and stepwise synthesis reactions are subject to de-phasing effects that restrict the length and fidelity of reads. In contrast, third generation sequencing aims to observe single DNA molecules in real time, avoiding amplification and enabling much longer reads. Two com-

mercially successful platforms exemplify this approach: Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT).

PacBio's technology is based on single-molecule real-time (SMRT) sequencing performed in zero-mode waveguides (ZMWs). These nanophotonic chambers enable the observation of individual DNA polymerases incorporating fluorescently labeled nucleotides [31]. Initially, the platform produced continuous long reads (CLR), which could exceed tens of kilobases in length. However, CLR reads had relatively high raw error rates, often above 10 percent, primarily due to insertions and missed events. While the errors were largely random and could be reduced by high coverage and consensus methods, they posed limitations for certain applications [117].

To address the need for higher accuracy, PacBio introduced circular consensus sequencing (CCS), which enables the generation of high-fidelity (HiFi) reads. In this method, a single circularised DNA molecule is sequenced multiple times by the polymerase, producing multiple subreads from which a consensus is calculated [140]. This trade-off reduces maximum read length compared to CLR, typically yielding HiFi reads in the range of 10 to 25 kilobases, but with accuracy exceeding 99.9 percent (Q30 or higher). HiFi reads combine the advantages of long-read context with the precision typically associated with short-read platforms, making them highly effective for applications such as variant detection, haplotype phasing, de novo assembly, and full-length transcript isoform analysis [140, 102].

HiFi sequencing has also become central to efforts aiming to generate high-quality, gapless reference genomes, such as those from the Telomere-to-Telomere (T2T) and Human Pangenome Reference Consortium (HPRC) projects [102, 108].

ONT sequencing operates using a different mechanism, based on the electrophoretic translocation of DNA through a protein nanopore embedded in a membrane. As the DNA strand moves through the pore, the nucleotide sequence influences the ionic current in a characteristic, though highly context-dependent, manner [67, 20]. This signal is decoded into nucleotide sequences using trained basecalling algorithms. Nanopore sequencing supports read lengths of hundreds of kilobases, with some individual reads exceeding one megabase [62]. Although early nanopore data were associated with high raw error rates (often greater than 15 percent), recent advances in basecalling algorithms, pore chemistry, and signal processing have led to substantial improvements in both read quality and yield [141]. The platform also offers the capacity for real-time, portable sequencing, with applications in clinical diagnostics, outbreak response, and environmental genomics [115].

Together, these third generation technologies enable a real time selection of reads that contain a desired sequence, access to structural variants, complex genomic regions, and full-length haplotypes that are difficult to resolve using short-read methods. Their capacity for

long, accurate reads is increasingly being used in hybrid sequencing strategies and standalone genome projects aiming to achieve chromosome-level completeness [94, 102].

These innovations have also been transformative in plant genomics, where genome size, ploidy, and high repeat content often hinder assembly using short reads alone. In recent years, long-read sequencing has enabled the generation of telomere-to-telomere assemblies in plants such as *Arabidopsis thaliana* [101], *Oryza sativa* (rice) [149], and *Zea mays* (maize) [77]. These assemblies have revealed large structural rearrangements, centromeric complexity, and previously missing gene content, contributing significantly to our understanding of genome architecture, evolution, and functional diversity in crops. In polyploid species such as wheat (*Triticum aestivum*) and canola (*Brassica napus*), high-fidelity long reads have been crucial for resolving homoeologous regions and assigning haplotypes across subgenomes.

Moreover, in pan-genome studies, long-read sequencing has helped identify presence–absence variation and structural polymorphisms in diverse collections, refining our understanding of gene content diversity, domestication, and adaptation. As these technologies continue to improve, they are expected to play a central role in both fundamental plant genomics and applied breeding.

1.6 Assemblies and reference genomes

The process of sequence assembly involves reconstructing longer contiguous sequences from overlapping fragments of DNA. This approach has a long history and was originally applied to smaller genomic regions, such as bacterial artificial chromosomes (BACs) or individual genes. As sequencing technologies have evolved, so have assembly strategies, shifting from clone-based and low-throughput methods to highly automated, high-resolution approaches. A central concept in this field is the reference genome, a curated genome assembly used as a standard for downstream analyses, including variant calling, gene annotation, and comparative genomics.

Due to technical limits that are unlikely to ever be fully eliminated, individual DNA sequence reads are rarely able to cover the entire genome of an organism. In an ideal scenario, a single sequencing read would span an entire DNA molecule, either a complete circular genome, as found in many prokaryotes, or an entire linear chromosome in eukaryotes. While this may become feasible in the future, current sequencing technologies still produce reads that are substantially shorter than these targets, even in the case of long-read platforms. This means that in most cases, the best sequencing data possible is a set of random reads sampled from fragments of the genome. In whole-genome “shotgun” sequencing, the genome is fragmented, often by sonication or enzymatic digestion, and the resulting fragments are

sequenced and reassembled computationally. This reconstructive step, known as assembly, has been an essential process since the first whole-genome sequence of bacteriophage X174 in 1977 [119].

Early sequencing efforts used primer walking, a targeted approach in which new sequencing reads were generated sequentially based on known regions of previously obtained sequences [1]. This method allowed for the gradual extension of sequences without relying on random fragmentation, and simplified the overlap-layout task during assembly. As genome projects scaled up, particularly for large and complex genomes, hierarchical shotgun sequencing became the preferred strategy. In this approach, the genome was first fragmented into large segments (typically 100–300 kb), which were cloned into bacterial artificial chromosomes (BACs). These BACs were physically mapped to determine their order and location before sequencing, and the resulting local assemblies were then combined into a full genome. This method, while time- and resource-intensive, enabled accurate assemblies for early reference genomes, including human (2001) [72, 136], rice (*Oryza sativa*, 2002)[44, 145], and *Arabidopsis thaliana*, the first plant genome (2000) [6].

During this time, overlap-layout-consensus (OLC) algorithms became the dominant assembly method. These algorithms work in three phases: first, establishing overlaps between reads (a computationally expensive problem with approximate $O(N^2)$ complexity), then arranging those overlaps into a consistent order (layout), and finally deriving a consensus sequence. Early OLC assemblers, such as TIGR [134], GigAssembler [68], and the Celera assembler [99], were critical to the success of landmark projects like the Human Genome Project. However, they often required substantial manual "finishing," especially in repetitive regions.

To reduce computational burden and improve tractability, some OLC methods introduced heuristic overlap detection and repeat masking strategies. The Celera assembler, for example, pioneered the integration of repeat masking and quality scoring into the assembly pipeline, setting a foundation that continues to influence modern long-read assemblers such as Canu [70] and FALCON [18].

The whole-genome shotgun (WGS) approach bypassed BAC-based mapping by directly fragmenting and sequencing the entire genome. With the advent of second-generation (short-read) sequencing, WGS became the dominant method [99, 11]. However, short reads (100–300 bp) from platforms like Illumina made it difficult to resolve repetitive regions, a challenge especially prominent in plant genomes. For instance, assembling the maize genome (*Zea mays*), which is over 80% repetitive, required paired-end reads, mate-pair libraries, and extensive scaffolding [121].

The repeat problem was also tackled through the use of de Bruijn graphs (DBGs), which fragment input reads into k-mers and connect them through k-1 overlaps. DBGs enabled memory-efficient assemblies for short-read datasets, and formed the core of tools like Velvet [146], ABySS [129], and SOAPdenovo [79]. They provided a scalable alternative to OLC, particularly for Illumina data. However, DBGs struggle with high-error data (like long reads) and reduce contiguity by discarding non-exact matches, an issue that modern assemblers try to mitigate through hybrid and graph-aware approaches [28].

Short-read assemblies often resulted in fragmented draft genomes, useful for gene-level studies but limited at the chromosomal scale. A key limitation of short-read sequencing is its inability to resolve repetitive regions, which leads to collapsed or misassembled sequences and gaps in assembly. To improve assembly continuity, scaffolding techniques such as optical mapping, Hi-C, and linkage mapping were used. These approaches were instrumental in the assembly of polyploid or large genomes such as wheat (*Triticum aestivum*), soybean (*Glycine max*), and grapevine (*Vitis vinifera*).

The emergence of third-generation sequencing technologies such as PacBio and ONT revolutionized genome assembly. Their long reads span repeats and structural variants, reducing the need for scaffolding and manual finishing. PacBio's HiFi reads, generated through circular consensus sequencing, combine high base accuracy with long-range continuity, ideal for complex plant genomes [140].

Assemblers like Canu [70] and FALCON [18] use the principles established in OLC and extend them to high-error, long-read data, relying on efficient string-graph constructions and repeat-aware heuristics. These tools have produced complete or near-complete genomes without human finishing, including tomato, barley, amaranth, and others [90, 55, 83, 89].

A reference genome is typically a high-quality assembly derived from a single individual or clone and serves as a coordinate system for annotation and variant analysis. However, no single genome captures the full genetic diversity of a species. This limitation is particularly acute in plants, where presence–absence variation, structural polymorphism, and transposable elements are common [54].

1.7 Resequencing

Given the high cost and complexity of assembling complete, error-free genomes for every individual, it is standard practice to use a high-quality genome assembly as a common coordinate system for comparative analysis. In resequencing studies, DNA from additional individuals of the same species is sequenced using shotgun sequencing libraries and then aligned directly to this reference genome. This approach enables the identification of genetic

variation without the need to perform *de novo* assembly for each sample. The process is known as resequencing, to distinguish it from whole genome sequencing and assembly. As an extension of this method, a hybrid approach has been developed in which reads that do not align to the reference are assembled separately and then anchored back to the reference genome [107].

Resequencing generally involves two analytical phases. In the first, sequencing reads from each individual or population are aligned to the reference genome using sequence alignment algorithms. In the second phase, these aligned reads are analysed locus by locus to identify allelic variants relative to the reference, including single nucleotide polymorphisms (SNPs), small insertions and deletions (indels), and other structural variants.

When applied to natural populations, resequencing allows for the systematic characterisation of genome-wide diversity. It provides insights into population structure, admixture, demographic history, and adaptive evolution. Genome-wide resequencing has been instrumental in uncovering patterns of selection, gene flow, and domestication in a variety of species. For example, resequencing datasets from wild and cultivated accessions of rice, maize, barley, and tomato have revealed loci associated with environmental adaptation and human-mediated selection.

Several large-scale population resequencing initiatives across diverse organisms have fundamentally shaped our understanding of genome-wide variation and evolution. In *Homo sapiens*, projects such as the 1000 Genomes Project [2] and the UK Biobank [14] have produced deeply sampled genomic datasets for tens to hundreds of thousands of individuals, providing insight into human genetic diversity, demographic history, and medically relevant variation. In *Saccharomyces cerevisiae*, population-scale studies have examined global collections of wild and industrial strains, illuminating domestication events, structural variation, and metabolic adaptation [111].

Comparable efforts in model and crop plants have similarly transformed evolutionary and functional genomics. Notably, the Arabidopsis 1001 Genomes Project provided whole-genome resequencing data for over 1,100 natural accessions, revealing detailed population structure, local adaptation, and extensive presence–absence variation [3]. The 3,000 Rice Genomes Project enabled the characterization of genomic diversity in a globally important crop, supporting trait mapping and the development of a rice pangenome [4]. Similarly, the 10+ Wheat Genomes Project has facilitated the analysis of complex polyploid genomes and expanded the available reference space [138].

Together, these initiatives demonstrate the power of resequencing to build population-scale genomic resources, enabling genome-wide association studies, pangenome devel-

opment, and comparative population genetics beyond the constraints of single-reference models.

1.8 Alignments and variant calling

Read alignment and variant calling are central steps in the analysis of resequencing data, and both are deeply rooted in algorithmic and probabilistic frameworks. During alignment, the goal is to determine the most likely genomic origin of each sequencing read with respect to a reference genome. Short-read aligners such as BWA [76] and Bowtie2 [73] employ compressed data structures based on the Burrows–Wheeler Transform (BWT) and the FM-index to perform fast and memory-efficient approximate string matching. These algorithms enable rapid location of candidate alignment sites while allowing for mismatches and small gaps, typically using a seed-and-extend approach. The alignment process results in a mapping quality score (MAPQ) that reflects the confidence in the placement of each read, taking into account uniqueness and potential alternative alignments. Reads mapping to repetitive regions or shared sequences, such as paralogous genes or transposons, often receive lower scores or are ambiguously mapped, which can impact downstream analyses [76].

In the context of long-read sequencing technologies, which exhibit higher per-base error rates but provide longer sequence context, aligners like Minimap2 have been developed. These tools use minimizer-based indexing, a strategy that selects representative k-mers within a read to reduce computational load while maintaining sensitivity to sequence similarity over large spans [75]. This method is particularly suited to noisy but structurally informative reads from technologies like Oxford Nanopore and PacBio.

Once reads are aligned, variant calling algorithms are applied to detect genomic differences between the sample and the reference. These differences can include single nucleotide polymorphisms (SNPs), small insertions and deletions (indels), and, in some cases, larger structural rearrangements. Variant callers operate by estimating the likelihood of each possible genotype at a given locus, given the aligned read data, sequencing error profiles, and prior assumptions about allele frequencies. Tools like GATK HaplotypeCaller [91] perform local *de novo* assembly within regions of interest to account for complex variation and realign reads to a set of candidate haplotypes. They also incorporate population-level information, such as known variant sites or joint genotyping across samples, to increase confidence in genotype calls. FreeBayes uses a similar haplotype-based Bayesian model but processes alleles jointly across individuals when used in population mode [40]. More recently, deep learning approaches such as DeepVariant have emerged, converting aligned reads into image-

like representations and using convolutional neural networks to infer genotypes directly from base-level data [113].

The output of these tools is typically stored in the Variant Call Format (VCF), which includes not only the position and identity of each variant but also associated metadata such as genotype likelihoods, read depth, allele balance, and predicted impact [24]. These annotations are critical for downstream filtering, prioritization, and interpretation. While this framework is robust for the detection of common variant types in well-mapped regions, it remains limited by its dependence on a single linear reference genome. Highly divergent haplotypes, structurally complex loci, or sequences absent from the reference may go undetected or be misrepresented. Such limitations have prompted the development of alternative models, including graph-based genome representations and multiple-reference strategies, which aim to improve variant discovery in structurally variable or highly polymorphic populations [110, 41].

1.9 Reference bias

While reference-based resequencing offers an efficient framework for variant discovery and comparative genomics, it also introduces a well-documented limitation known as *reference bias* [25]. This bias arises from the reliance on a single reference genome during read alignment and variant calling. Because the reference represents only one haplotype, often derived from a single individual or a composite of few individuals, it fails to capture the full spectrum of genetic diversity present within a population or species [9].

Reference bias affects both the alignment and variant detection phases. Most short-read aligners attempt to find the best possible match of a sequencing read to a linear reference genome. However, true biological variants, particularly insertions, deletions, or rearrangements that are not represented in the reference, can significantly reduce the mapping quality or even cause reads to fail to align [133]. In such cases, the reference allele may be preferentially observed, while the non-reference variant may be underrepresented or misinterpreted. This effect extends even to single nucleotide polymorphisms (SNPs), where subtle biases in read alignment can skew allele balance estimates [132].

The consequences of reference bias are particularly pronounced in structurally complex or highly polymorphic regions, and in individuals or populations that are genetically distant from the reference genome. In downstream analyses, this can lead to systematic underestimation of nucleotide diversity, misestimation of population structure, and distortion of selection signals [12]. In experimental contexts that rely on accurate allele, these distortions can introduce serious artifacts [148].

In plant species, which often contain extensive structural variation and complex genome architectures, reference bias is particularly relevant. In *Arabidopsis thaliana* for example, the reference genome TAIR10 is based on the Col-0 accession, which is genetically distant from many globally distributed natural accessions. Studies such as the 1001 Genomes Project have shown that accessions more closely related to Col-0 tend to yield more confident alignments and fewer detected variants, whereas divergent accessions like Cvi-0 or Sha exhibit increased rates of unmapped reads, lower variant sensitivity, and inflated reference allele calls [3, 64]. This issue is especially pronounced in structurally variable regions, transposable element-rich intervals, and gene presence–absence polymorphisms common in plants [45].

Although improvements in sequencing technology, particularly long-read platforms, can help resolve some types of reference bias by spanning repetitive or structurally variable regions, they do not eliminate the fundamental problem. As long as reads are interpreted relative to a single linear reference, any variant or haplotype absent from that reference will be incompletely represented [15]. Furthermore, the cost and throughput advantages of short-read sequencing continue to make it the dominant approach in large-scale population resequencing, where reference bias remains a persistent concern. Particularly in population-scale projects using short reads, such as in many plant breeding or diversity studies, reference bias continues to pose a major challenge for accurate variant discovery and downstream interpretation.

1.10 Pangenomes

The pangenome can be defined as the complete collection of genomic sequences present within a species, encompassing both core sequences shared by all individuals and accessory or variable sequences found only in subsets of the population. It represents an idealized view of a species' full genetic diversity, rather than the genomes of a small or arbitrary set of individuals. As such, the pangenome serves as a conceptual framework for studying intraspecies variation, and has motivated the development of various formal models for its representation.

The increasing recognition that single reference genomes fail to capture the full extent of genetic variation within a species has led to the development of pangenome representations [9, 45]. These aim to model not only the shared (core) genome, but also the variable (accessory) and unique sequences found among individuals. Several strategies have been proposed to represent pangenomes, ranging from simple linear models to complex graph-based structures that support multi-genome analysis and alignment.

In early pangenome studies, variation was often described in terms of presence–absence matrices or reference-anchored alignments [135]. In this approach, genomic or gene-level sequences from multiple individuals are aligned to a single linear reference genome, and differences are recorded in tabular form. While these models are easy to interpret and integrate with conventional analysis pipelines, they introduce reference bias and do not allow for the direct representation of complex structural variants or rearranged haplotypes. Additionally, reads that originate from sequences absent in the reference may not align accurately or at all, leading to an underestimation of diversity.

To address these limitations, graph-based pangenome models have become increasingly common [110, 41]. In graph representations, genomes are encoded as networks of nodes and edges, where each node represents a DNA sequence segment, and edges describe the allowed connections between sequences. This allows for multiple sequence paths through the graph, each corresponding to a different haplotype or genome. Such models naturally incorporate single nucleotide variants, insertions and deletions, inversions, duplications, and presence–absence variation within a unified framework.

Several types of graphs are used in practice. Sequence graphs, often constructed from whole-genome alignments, represent contigs or chromosomes from multiple genomes as paths through a graph. These models are implemented in tools such as Minigraph-Cactus [78, 53] and PGGB [39], which support efficient indexing, navigation, and visualization of multi-genome structures. Variation graphs, as used in the VG toolkit [41], are designed to represent small and large variants in a directed graph structure, enabling accurate read mapping and variant calling across diverse haplotypes. Another approach, rooted in genome assembly, uses de Bruijn graphs to represent shared and divergent k-mer content across genomes. While computationally efficient and scalable, de Bruijn graphs are less suited for the analysis of long-range structural variants and may be more difficult to interpret biologically. Colored graphs, which annotate each node with information on which genomes contain the sequence, provide a way to track core and accessory elements across individuals or populations [59].

Graph-based pangenomes offer several advantages over linear references. They reduce reference bias in read mapping, improve variant detection in structurally complex regions, and enable comparative analysis across a wider range of genotypes [130, 41]. They are particularly powerful in plant species, where high levels of structural variation and gene presence–absence are common [63]. However, these models also present challenges. Graph construction and manipulation require significant computational resources, and the majority of bioinformatic tools and file formats are still designed around linear references. Moreover,

new strategies are needed to annotate and interpret biological features within graph structures, including genes and regulatory elements.

1.11 Applications with pangenomes

Pangenomes have become indispensable resources in genomics, offering a more comprehensive and unbiased representation of genetic diversity compared to traditional linear reference genomes. By incorporating polymorphisms, structural variants (SVs), and alternative haplotypes into a unified graph structure, pangenomes, especially those encoded as variation graphs, enable a wide array of genomic analyses.

Traditional short-read aligners, such as BWA and Bowtie2, map sequencing reads to a linear reference genome, which can result in mismatches or alignment failures in regions with structural variants or divergent haplotypes. Graph-based mappers, including `VG map`, `Giraffe`, and `GraphAligner`, allow reads to align against all known variant paths simultaneously. This approach reduces reference bias and improves alignment accuracy, particularly in genetically diverse populations or species with complex genomic architectures [41, 130, 116].

Once reads are mapped to a variation graph, variant calling can be performed using graph-aware tools. The VG toolkit supports genotyping by traversing paths in the variation graph and estimating the likelihood of alternative alleles. This method enhances the detection of SNPs, indels, and large structural variants, which are often missed by linear reference-based methods [38].

Pangenomes facilitate genome inference and genotype imputation across populations. Tools like Graphtyper and PanGenie leverage graph-based haplotype panels to impute missing genotypes, improving accuracy over linear reference panels, especially in underrepresented populations [29, 30].

Variation graphs allow for fine-scale comparison of multiple genomes by encoding multiple assemblies or haplotypes as paths within the graph. This capability enables the efficient identification of conserved and divergent regions, which is particularly valuable in species like maize, wheat, and brassicas, where traditional pairwise alignments may not capture the full extent of genomic diversity and rearrangement patterns [110, 32].

Pangenomes serve as foundational structures for building graph-based indexes, facilitating rapid retrieval of genomic features and variants across samples. Tools such as ODGI support visualization and navigation of these graphs, aiding in the generation of reference coordinate systems that remain consistent across versions and assemblies. This consistency enhances reproducibility, portability, and scalability in genomic workflows [46].

The adoption of pangenomes, particularly variation graph-based models, is transforming the representation, analysis, and interpretation of genomic diversity. By integrating the full spectrum of variation directly into the reference structure, these models reduce biases, improve accuracy, and enable more inclusive and representative studies in both basic and applied genomics.

1.12 Variation graph construction

The construction of pangenome graphs is a critical step in representing the full (known) spectrum of genomic variation within a species. Among the most prominent tools for this purpose are PGGB (Pangenome Graph Builder) [39] and Minigraph-Cactus (MC) [53], each represent distinct philosophies and technical approaches. Both aim to encode genome-scale variation into graph structures that can be used for read mapping, variant calling, and comparative analysis, but they differ significantly in methodology and applications.

Minigraph-Cactus is a hybrid framework developed by the Human Pangenome Reference Consortium (HPRC) to efficiently build reference-based pangenomes from large numbers of high-quality assemblies [53]. First, Minigraph constructs a coarse variation graph by aligning new genome assemblies against a linear reference backbone using minimizer-based indexing [80]. This results in a draft graph capturing major structural differences such as insertions, deletions, and duplications. In the second phase, the Cactus aligner refines the graph by performing reference-free multiple whole-genome alignment [7]. This step captures more complex and nested structural variants, including inversions and rearrangements, and improves the accuracy of the graph structure. The resulting graph retains a reference-like coordinate space while supporting accurate variant representation across diverse haplotypes. Minigraph-Cactus has been used to generate human reference pangenomes and is particularly well suited for scaling to hundreds of assemblies with relatively conserved synteny.

In contrast, PGGB (Pangenome Graph Builder) offers a fully reference-free approach [39]. Rather than anchoring to a linear reference genome, PGGB aligns all genome sequences to each other using homology-based methods and constructs a variation graph from these dense alignments [32]. The workflow involves three major components: `wfmash` for pairwise whole-genome alignment, `seqwish` for graph induction from the alignments, and `smoothxg` for graph normalization and path sorting. The final product is a highly detailed graph where all sequence variation is represented symmetrically, and no single genome is privileged as a coordinate system. This makes PGGB particularly useful in contexts where structural divergence is high, or where a reliable reference genome is lacking. However, the computational

demands of this approach are significantly greater than those of Minigraph-Cactus, especially for large datasets.

The choice between Minigraph-Cactus and PGGB depends on the nature of the study system and analytical goals. Minigraph-Cactus is favored for large-scale projects involving closely related genomes and requiring scalability, while PGGB excels in capturing fine-grained variation in highly diverse or poorly referenced species. Both methods contribute to the growing toolkit of graph-based genome analysis and represent important steps toward comprehensive and unbiased representations of genomic diversity.

1.13 Overview and Objectives

Genome graphs offer a more comprehensive representation of genomic variation than linear reference genomes by integrating multiple assemblies into a single, unified structure. They allow for the inclusion of structural variants and population-level diversity, making them increasingly relevant for comparative and population genomics. However, despite their potential, variation graphs remain difficult to analyze and interpret in practice. Current methods are often limited in flexibility, scalability, or biological interpretability, particularly when applied to large and complex datasets.

This thesis addresses these challenges by developing new tools and methods for analyzing and interpreting genome graphs. A central focus is the implementation of `gretl`, a statistical framework designed to extract structural and path-based metrics from variation graphs. This enables detailed comparison of graph construction parameters and methods. In addition, the tool `gfa2bin` was developed to support graph-based genome-wide association studies (GWAS) using node-level coverage derived from short-read alignments. To further enable the study of variation, a novel algorithm was introduced that detects bifurcation-based structures, offering an alternative to classical bubble models.

All methods were applied to real-world datasets, with a focus on the 1001 Genomes *Arabidopsis thaliana* collection. The overall objective of this thesis is to improve the usability, scalability, and biological interpretability of genome graphs through modular, graph-native analysis tools.

Chapter 2

Evaluating variation graphs

This chapter is based on the publication *gretl – Variation Graph Evaluation TooLkit*, available on *Bioinformatics - Application notes* at https://academic.oup.com/bioinformatics/article/41/1/btae755/7932228.

2.1 `Gretl` - Graph evaluation toolkit

Advances in short-read based resequencing have greatly improved our understanding of genomic variation in many different species ([3, 111, 2]. More recently, long reads have made it possible to assemble complete genomes with remarkable speed and precision. Moving from variant inventories to complete genomes facilitates much more comprehensive analysis and genome-wide comparison between samples. As an example, in the plant *A. thaliana*, the level of detail provided by (nearly) complete genomes has already led to new insights into conservation of synteny and to much more accurate description of single nucleotide polymorphisms (SNPs), copy number variants (CNVs), and structural rearrangements [65, 43].

To mitigate the biases associated with a single reference genome, pangenomes built from diverse sample collections are being created from increasingly complex genomes [81, 125]. A crucial tool for efficient storage and comprehensive analysis of genetic variations within diverse and intricate genomic regions is the variation graph, which condenses similar sequences into nodes and captures variations in a reference-free manner. Graph shape and structure depend on the choice of construction method and parameter set, requiring tuning and adjustment based on the genome complexity [74] and the research question, highlighting the need for a comprehensive evaluation tool.

Genome graphs are typically stored in GFA (Graphical Fragment Assembly) format, a

standardized data format, which is also the main input for `gretl`, the tool introduced here. Nodes in the graph represent DNA segments, connected by edges and each node has an associated DNA sequence and a unique identifier. GFA can store additional information like allele frequency, quality scores, or annotations, if needed. The format ensures interoperability among software tools, facilitating collaboration and analysis development. `Gretl` fully supports GFAv1 files[1], ensuring interoperability across a wide range of graph tools. Adopting GFAv2 is an option for the future, as more upstream graph tools migrate to GFAv2.
Several tools for genome graph analysis are available and being actively developed, including odgi [46], vg [41] and gfastats [35]. While odgi and vg offer powerful platforms for modifying and analyzing genome graphs, there is still a need for tools that can rapidly compute an overview of a large number of statistical features for evaluation of variation graphs. Although gfastats is designed for statistics, its primary focus lies in assembly graphs, which have, in comparison with whole-genome graphs, distinct characteristics. While it does provide several useful statistics for genome graphs, its main function remains an overall toolkit for modifying GFA files and delivering high quality single individual genomes. In our benchmarking and comparison between the different methods, we excluded gfastats, because the run did not finish within a reasonable amount of time. One of the primary motivations behind our work was to provide a fast and efficient tool for the initial evaluation of newly constructed graphs. Building genome graphs is a complex process, and one often needs to rapidly assess their quality. `gretl` aims to address this need by offering an all-in-one tool that evaluates graph structure and composition. With `gretl`, researchers can evaluate the graph using a variety of quantitative metrics and identify potential areas that require further investigation or refinement. As an example, graphs with high average depth are most likely highly collapsed, merging duplicated segments such as TEs into a single structure, which in turn makes it harder to align sequences or sequencing reads to graphs, but it helps to understand the nature of transposed DNA segments or copy-number variation (CNV). Moreover, one can generate statistics on genome-growth, pangenome distribution and/or for specific paths, a feature used in the subsequent analyses in this work (Figure 3.6). This preliminary information can guide subsequent analyses and describe properties of different speciesṕangenomes, providing a solid foundation for further investigations.

`gretl` offers valuable insights into genome graphs constructed using PGGB [39] and Minigraph-Cactus [53], as well as other graphs in GFA format. The only requirement is the availability of numeric node IDs, which can, if not already present, be converted from non-numeric node IDs via the `gretl node2int` subcommand. `gretl` provides several subcommands that offer comprehensive graph analysis, covering aspects such as graph

[1]http://gfa-spec.github.io/GFA-spec/GFA1.html

complexity, interconnectedness, and node degree. We provide Python scripts and follow-along markdowns that can be used for post-processing and visualization of the output similar to the plots shown in the figures here, allowing further exploration and interpretation of the results.

2.2 Materials and Methods

2.2.1 Materials

Arabidopsis thaliana genomes were from [143], while *S. cerevisiae* genomes were from [103]. An available PGGB-built human genome graph [81][2] was used. Chromosomes 14, 18, 19, 21, 22 were downloaded from the AWS HPRC repository[3].

2.2.2 Methods

Graph construction

To illustrate the capabilities of `gretl` for graph evaluation, we used pangenome graphs constructed from chromosomes of *H. sapiens* (n=48 for Chr 14, 18, 19, 21, 22), *S. cerevisiae* (n=30 for Chr 1, 3, 5, 9, 10) and *A. thaliana* (n=67 for Chr 1-5). Details about the datasets [81, 103, 143] and graph construction are given in Table 2.1 and Table 2.2.

The *A. thaliana* and *S. cerevisiae* graphs were constructed using the PGGB pipeline [39] for each chromosome individually. We used the following parameter to construct the *S. cerevisiae* graph for the species comparison experiment: -p 90; -k 31, -n 30 -s 5000 -asm10. *Arabidopsis thaliana* graphs were constructed with the following parameters: -p 90 -s 10000 -G 2000 -n 67 -t 32 -k 49 -P asm5 -O 0.001 -G 700,900,1000. Parameter comparisons were conducted using *S. cerevisiae* genomes with various parameter combinations. The parameters we modified included -s (2k, 5k, 10k), -p (80, 90, 95), -k (19, 31), -n (15, 30, 60), and -P (asm5, asm20). The PGGB workflow was executed with wfmash (v0.10.2-2-gb310bd1), seqwish (v0.7.8-3-gd9e7ab5), odgi (v0.8.2-92-gbfae0b3), and smoothxg (v0.6.8-31-g06bbf35).

We used BandageNG[4] (version: v2022.09), a fork of the original Bandage[5], for visualization of variation graphs.

[2]https://github.com/human-pangenomics/hpp_pangenome_resources
[3]https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=pangenomes/freeze/freeze1/pggb/chroms/
[4]https://github.com/asl/BandageNG
[5]https://github.com/rrwick/Bandage

Table 2.1 **Data sets used in our experiments.** *Arabidopsis thaliana*[6] and *Saccharomyces cerevisiae*[7] are telomere-to-telomere assemblies, wheras the human[8] genomes consist of multiple contigs. For *A. thaliana* and *S. cerevisiae*, we utilized the genome to construct new graphs for several experiments.

	A. thaliana	*S. cerevisiae*	*Homo sapiens*
Paths	66	30	1,072-3,029
Samples	66	30	49
Ploidy	1n	1n	2n
Chrosmosomes	5	16	23
Reference length [Mbp]	130	12	3,200

Table 2.2 `gretl` **- Information on the genome graphs**.
*Parameter set -p 90, -n 30, -s 5000, -k 31. **Parameter set: -p 90 -s 10000 -G 2000 -n 67 -t 32 -k 49 -P asm5 -O 0.001 -G 700,900,1000.

Organism	**Chr**	**Sequences**	**Samples**	**Nodes [x1000]**	**Edges [x1000]**	**Average node size [bp]**
Homo sapiens	**14**	1,882	48	4,155	5,790	65.7
	18	1,270	48	2,832	3,980	86.2
	19	1,072	48	3,021	4,215	96.2
	21	3,029	48	2,761	3,883	99.2
	22	1,757	48	3,760	5,224	123.4
Saccharomyces cerevisiae*	**I**	30	30	52	76	9.3
	III	30	30	53	73	11.4
	V	30	30	125	171	15.9
	IX	30	30	67	93	11.8
	X	30	30	74	101	11.7
Arabidopsis thaliana	**1**	67	67	6,891	9,741	17.0
	2	67	67	4,927	6,979	15.5
	3	67	67	5,977	8,520	15.4
	4	67	67	4,747	6,789	18.2
	5	67	67	5,657	7,984	19.6

Table 2.3 **Parameter sets used in the grid earch experiment.**
For additional details, please refer to https://github.com/pangenome/PGGB.

Setting	Meaning	Value
p	percent identity in the wfmash step	80
		90
		95
k	percent identity in the wfmash step	19
		31
n	number of mappings to retain for each segment	15
		30
		60
s	segment length for mapping	2000
		5000
		10000
asm	smoothxg alignment parameters (sequence divergence between sequences)	asm5 asm20

Parameter studies

For the comparison between different parameters, we used the following parameter combinations using 30 genomes of *Saccharomyces cerevisiae*.

Correlation Analysis

We evaluated the linear relationship between the parameter p and other numerical features using Pearson correlation. Correlations were computed with `scipy.stats.pearsonr`, and only features with P values below 0.05 were considered significant. Significant features were separated into positively and negatively correlated groups, sorted by correlation strength, and visualized using a horizontal bar plot.

2.2.3 Implementation

`gretl` has been implemented in the Rust programming language. It incorporates several Rust crates to enhance performance and enable multithreading. Furthermore, a new GFA format reader is provided as a library. In addition to handling the GFAv1 format utilized by `gretl`, this reader can interpret GFAv2 as well and it can be used by other tools.

Relationship between samples and path/walks

Genomes can contain multiple chromosomes. Depending on genome complexity and size, individual chromosomes may be represented in whole-genome assemblies as single or multiple contigs. To link multiple contigs and/or chromosomes to the same sample, we use PanSN-spec[9]. Similar to GFA walks, sample names are separated by haplotype ID and contig or scaffold name. Walks and paths are interchangeable using PanSN-spec. We work around these samples, which can be thought of as collections of multiple paths or walks drawn from the same assembly. The number of paths in a single graph is therefore at least as large as the number of samples.

Comparison with other tools

We compared `gretl` with vg and odgi. We ran vg version v1.54.0 "Parafada" and odgi version v0.8 (commit: v0.8.2-92-gbfae0b3). VG was run with the following flags: -z -l -L -s -H -T. We decided to exclude additional flags, since they report features of the graph that might be out of scope for a statistical view of the graph. Detecting and reporting bubbles (variation) of the graph is important, but it does cover an additional layer of the graph, which is out of scope for a fast and accessible statistical check. For odgi, we used the "-m" flag, which internally runs (-S, -W, -L, -b, -l, -g, -s, -f, -d, -p, -N).

Pangenome classification

We utilized the characteristics of the graph to classify different levels of relatedness for *S. cerevisiae* genomes. Nodes present in all accessions were annotated as core, nodes that were only traversed by one accession as private, and all other nodes (>1 and <30 traversals) were classified as soft (shell).

2.2.4 Evaluation

`gretl` facilitates in-depth comparison of specific graphs using a wide range of metrics. This analysis can be performed at both the graph level and the path level, providing researchers with comprehensive insights.

More detailed information for some statistics

Similarity/depth We define similarity by the number of samples traversing a single node. Depth counts the total amount of traversals, regardless of whether it is traversed by a single

[9]https://github.com/pangenome/PanSN-spec

sample multiple times or different samples. Similarity and depth can be normalized by sample number.

Jumps Links or edges are jumps from one node to another node. Where node IDs are consecutive integers, we can calculate the difference (in node ID) between two nodes. Large differences normally reflect a link/edge that is not in pan-genomic order and therefore represents an indel. If the node IDs in the path are linearly increasing, there is a high probability that the graph structure is linear too. Node IDs are not required to be in consecutive order, except for jump related statistics, which require a sorted graph. To this end, the command "odgi sort" can be used, which sorts node IDs in pangenomic order[10].

Node degree Node degree defines the number of edges linking to a single node. We calculate this statistic for all nodes in the graph and compute average, median and standard deviation. Since links/edges are directed in GFA files, we also report "incoming" and "outgoing" node degree separately.

Graph-centric

Graph-centric statistics refer to metrics that can be computed independently of specific path information within the graph. These include general structural properties such as node degree distributions, graph density, and connectivity measures. These features are generally not specific to sequence graphs and are often well-established in general graph theory.

[10]Learn more about sorting variation graphs here: https://academic.oup.com/bioinformatics/article/40/7/btae363/7705520

Table 2.4 **Graph-centric statistics reported by `gretl` *stats*.**
Types of values and metric types reported. Green: Reported by other tools. Orange: Can be computed by math operations. Red: Not reported by other tools.

Name	Metric type	Unit	Range of values	Description	vg	odgi
Paths	Single integer	Dimensionless	All positive	Number of P and W lines in the file		
Samples	Single integer	Dimensionless	All positive	Number of samples. Samples are collections of P/W-lines that are defined by PanSN-spec		
Nodes	Single integer	Dimensionless	All positive	Number of nodes		
Edges	Single integer	Dimensionless	All positive	Number of edges		
N/E ratio	Single number	Dimensionless	All positive	Nodes divided by edges		
Graph size	Single integer	bp	All positive	Total amount of sequence in the graph (um of all nodes) in bp		
Input genome size	Single integer	bp	All positive	Sum of sizes of all input genomes of the graph		
Compression	Single number	Dimensionless	All positive	Graph size divided by input genome size		
Node length	Average, median	bp	All positive	Node length in bp		
Node length top 5%	Average, median	bp	All positive	Average node length of the top 5 % nodes (sorted by size) in bp		
Bin	Single integer	Dimensionless	All positive	Number of nodes in each bin. Bin can be modified by user input		
Similarity	Average, median, std	Dimensionless	All positive	Average similarity of the entire graph		
Similarity (normalized)	Average, median, std	Dimensionless	0-1	Similarity divided by number of samples		
Depth	Average, median, std	Dimensionless	All positive	Average depth of the whole graph		
Depth (normalized)	Average, median, std	Dimensionless	0-1	Depth divided by number of samples		
Node degree	Average, median, std	Dimensionless	All positive	Average node degree. Average number of edges linking to one node (total).		
Inverted edges (normalized)	Average, median	Dimensionless	0-1	Number of edges that change their direction (from + to - or from - to +). Normalized by the total number of edges.		
Negative edges (normalized)	Average, median	Dimensionless	0-1	Number of negative edges. Here, a negative edge is defined to be from “-” to “-”. Normalized by the total number of edges.		
Self edges (normalized)	Raw, normalized	Dimensionless	0-1	Number of self edges. Self edges are edges starting and ending at the same node. Normalized by the total number of edges.		
Graph density	Single number	Dimensionless	All positive	Proportion of observed edges and nodes relative to the number of all possible edges. Calculation: Edges / ((Nodes * (Nodes - 1))/2)		

Path-centric

Path-centric statistics are derived from the traversal paths embedded within a graph structure. In this context, they are based on the *path* (P) and *walk* (W) lines as defined in the GFA format. These statistics are particularly informative when the graph has been constructed from whole-genome assemblies or when a high-quality pangenome representation (PanSN) is available. Path-centric features encompass both metrics specific to individual paths and statistics computed on the subgraphs induced by the traversal of a particular path or walk. Conceptually, they represent graph-centric statistics constrained to the subset of the graph visited by a given path.

Hybrid statistics

Any metric calculated at the path level can be summarized comprehensively across all paths in the graph. For example, we take path lengths and calculate the average and standard deviation graph-wide. The average indicates typical path length while standard deviation shows variability, revealing if paths are relatively consistent or contain outliers. High standard deviation may warrant further path-specific investigation. Where applicable, we also report total and maximum values for a metric across all paths to provide additional insights.

2.2.5 Analysis

The analysis of a specific graph is a crucial step for extracting information from a chosen graph structure. We provide a range of advanced analysis methods for individual graphs, including pangenome classification, complexity analysis using sliding windows, and bootstrapping-based approaches.

Sliding window analysis

For each path, we implement a node- or sequence based sliding window approach to highlight graph regions exhibiting extreme characteristics, whether highly divergent areas or large structural variations. The analysis can incorporate similarity, node size, depth, or other metrics within each window. Visualizing the sliding window data as a heatmap reveals path regions arising from the same graph loci and clarifies differences.

Bootstrapping

We have implemented a bootstrapping approach to estimate genome graph growth when adding more samples. Our method takes the full graph and randomly removes subsets of

Table 2.5 **Path-centric statistics reported by `gretl` *stats*.**
Descriptions for each value and which metric type is reported when run in the default mode. Each value is calculated for the individual path independently and then summarized by the reported metric. Using the "-p" flag reports the independent values for each path separately.

Name	Metric type	Unit	Range of values	Description	vg	odgi
Sequence	average, std	bp	All positive	Total amount of sequence in the path		
Covered	average, std	Dimensionless	All positive	Sequence [bp] / Graph size [bp]		
Nodes	average, std	Dimensionless	All positive	Number of nodes in the path		
Unique edges	average, std	Dimensionless	All positive	Number of unique nodes		
Directed nodes	average, std	Dimensionless	All positive	Number of different directed nodes		
Edges	average, std	Dimensionless	All positive	Number of edges		
Unique Edges	average, std	Dimensionless	All positive	Number of unique edges		
Unique nodes	average, std	Dimensionless	All positive	Number of unique nodes		
Unique nodes	average, std	bp	All positive	Amount of sequence of unique nodes		
Unique nodes (normalized)	average, std	Dimensionless	0-1	Unique nodes / Nodes		
Unique nodes (normalized)	average, std	Dimensionless	0-1	Unique nodes [bp] / Sequence [bp]		
Unique edges (normalized)	average, std	Dimensionless	0-1	Unique Edges / Edges		
Inverted nodes	average, std	Dimensionless	All positive	Number of inverted nodes		
Inverted nodes	average, std	bp	All positive	Amount of sequence of all inverted nodes		
Inverted nodes (normalized)	average, std	Dimensionless	0-1	Inverted nodes / Nodes		
Inverted nodes (normalized)	average, std	Dimensionless	0-1	Inverted nodes [bp] / Sequence [bp]		
Jumps total	average, std	Dimensionless	All positive	Total number of of jumps		
Jumps total (normalized)	average, std	Dimensionless	All positive	Jumps total / Edges		
Jumps larger than X	average, std	Dimensionless	All positive	Number of jumps larger than X		
Jumps larger than X (normalized)	average, std	Dimensionless	0-1	Jumps larger than X / Edges		
Node size average	average, std	Dimensionless	All positive	Average node size		
Node size median	average, std	Dimensionless	All positive	Median node size		
Node size std	average, std	Dimensionless	All positive	Standard deviation of node size		
Depth average	average, std	Dimensionless	All positive	Average depth		
Depth median	average, std	Dimensionless	All positive	Median depth		
Depth std	average, std	Dimensionless	All positive	Standard deviation of depth		
Depth average (normalized)	average, std	Dimensionless	All positive	Depth average / Number of samples		
Depth median (normalized)	average, std	Dimensionless	All positive	Depth median / Number of samples		
Depth std (normalized)	average, std	Dimensionless	All positive	Depth std / Number of samples		
Similarity average	average, std	Dimensionless	All positive	Average similarity		
Similarity median	average, std	Dimensionless	All positive	Median similarity		
Similarity std	average, std	Dimensionless	All positive	Standard deviation of similarity		
Similarity average (normalized)	average, std	Dimensionless	0-1	Similarity average/ Number of samples		
Similarity median (normalized)	average, std	Dimensionless	0-1	Similarity median / Number of samples		
Similarity std (normalized)	average, std	Dimensionless	0-1	Similarity std / Number of samples		
Degree average	average, std	Dimensionless	All positive	Average node degree		
Degree median	average, std	Dimensionless	All positive	Median node degree		

samples, recalculating statistics on the reduced graphs. This bootstrapping process is repeated N times with different random sample subsets to assess graph metrics across a range of sizes. While computationally intensive, this approach accounts for effects of extreme outlier samples unlike faster approximations.

Core-Pan analysis

Of general interest is the analysis of pan-core-private parts of the genome in regards on whole populations. Similar to allele frequencies, we can calculate for each node the number of samples covering this node. From that statistics we can classify the nodes into pan-core and private parts of the graph.

Feature list and adding new statistics

`Gretl` can be utilized to report several statistics within a features list – for each node, edge, or directed node, dependent on user input. Since `gretl` is open-source, it can can be modified and extended with new measures by forking the code on GitHub and submitting a pull request.

Complexity

Node or edge complexity was measured by analyzing either the entirety of nodes in the graph or a selected subset, depending on user input. For each node, we performed a breadth-first traversal up to N steps away (where N is user-defined), and accumulated statistics such as the number of nodes visited, the total sequence amount, or the distance from the original node. These values can then be visualized using scatter plots to highlight regions of interest, such as areas of high complexity, indicated by a dense concentration of nodes.

2.3 Results

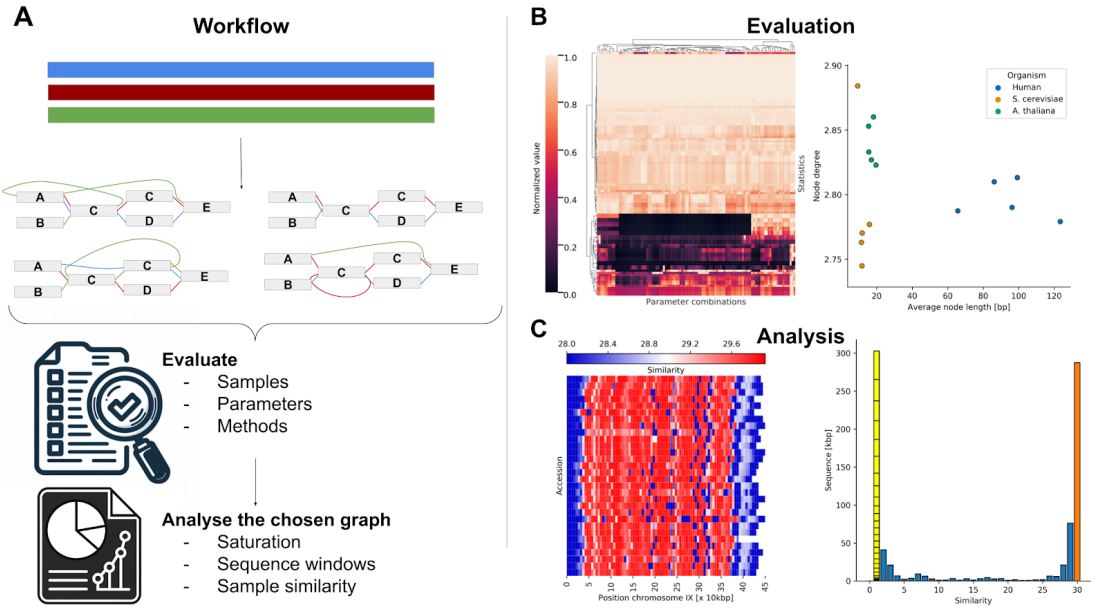


Fig. 2.1 `Gretl` **overview.**
(A) Genome graph construction workflow: Genome graph properties are influenced by various factors, including parameter selection, sample curation, and methodology, all of which impact the layout and structure of the resulting genome graph. For evaluation purposes, multiple graphs can be simultaneously generated and compared to identify an optimal representation for a specific task. The selected graph can then be analyzed with `gretl`. **(B)** Visualization of `gretl` output: Left, graphs can be clustered based on multiple statistics, grouping similar species or construction parameters (shown here, with normalized values). Right, scatter plot depicting two selected statistics across various graphs, facilitating comparisons between different species. (C) In-depth analysis of a selected genome graph (example from yeast): Left, path-centric sliding window analysis of the *S. cerevisiae* genome graph, highlighting regions of high similarity. Right, pan-genomic analysis of the genome graph. Sequences found only in a single sample are separated and each block represents one path of the graph.

The provided statistics enable evaluation of graphs built with different parameters from the same dataset (Figure 2.1B). Statistics generated by `gretl` (Figure 2.1) can guide subsequent analyses and describe the properties of different species' pangenomes.

Additionally, these values facilitate the evaluation and comparison of graphs from different methods or organisms, enabling insights into the complexity and structure of the genome (Table 2.2, Figure 2.8). It is important to note that some of these statistics may exhibit similar behavior and display high correlation due to their interconnected nature (Figure 2.11). Researchers can explore the impact of varying parameters during graph construction on the

same dataset or analyze different species by comparing or clustering their genome graphs by statistical features.

In general, `gretl` offers more comprehensive information about the graph than other tools. The tabular output format provides an easy overview as well as seamless integration with scripting languages such as R and Python for post-processing.

`gretl` facilitates in-depth comparison of specific graphs using a wide range of metrics. This analysis can be performed at both the graph level and the path level, providing researchers with comprehensive insights. At the graph level, various metrics and statistics can be explored to identify regions of interest, which can be further investigated in subsequent studies (Figure 2.1C, 2.12, 2.13). Sliding window analyses on sequence or node level give powerful insight into local complexity or distant sequence similarities (Figure 2.10). This could be, for example, useful to demonstrate the local complexity of possible QTL hits [96].

The path-centric analyses allow for computation of independent statistics and metrics for specific paths within the graph (Table 2.5). This approach enables comparisons between different samples or populations, helping in the identification of path-specific differences within the pangenome `gretl` - Variation Graph Evaluation TooLkit (Figure 2.1C, Figure 2.9). Furthermore, it enables the identification of samples that display isolated or otherwise distinctive representations in the graph (Figure 2.6). By carefully examining the paths within the graph, researchers can uncover structural patterns, variations, and potential functional significance embedded within the genome. This comprehensive analysis of paths should contribute to a deeper understanding of the genome's complexities and provide valuable insights for further research. A table of reported statistics including the name, description and availability in other tools can be found in Tables 2.4 and 2.5.

2.3.1 Benchmark

During testing on a 3.2 GHz AMD Epyc 64 core machine using chromosome 19 of the PGGB-built HPRC graph [81], which consisted of 48 samples (96 haplotypes, 1,072 paths) with 3.02 million nodes and 4.21 million edges, our evaluation tool demonstrated a level of performance that should greatly encourage its adoption for any genome graph construction workflow. It computed simple summary statistics from GFA files twice as fast as other approaches in under five minutes, utilizing 2.91 GB of memory (Figure 2.3). We observed almost linear scaling properties (Figure 2.2).

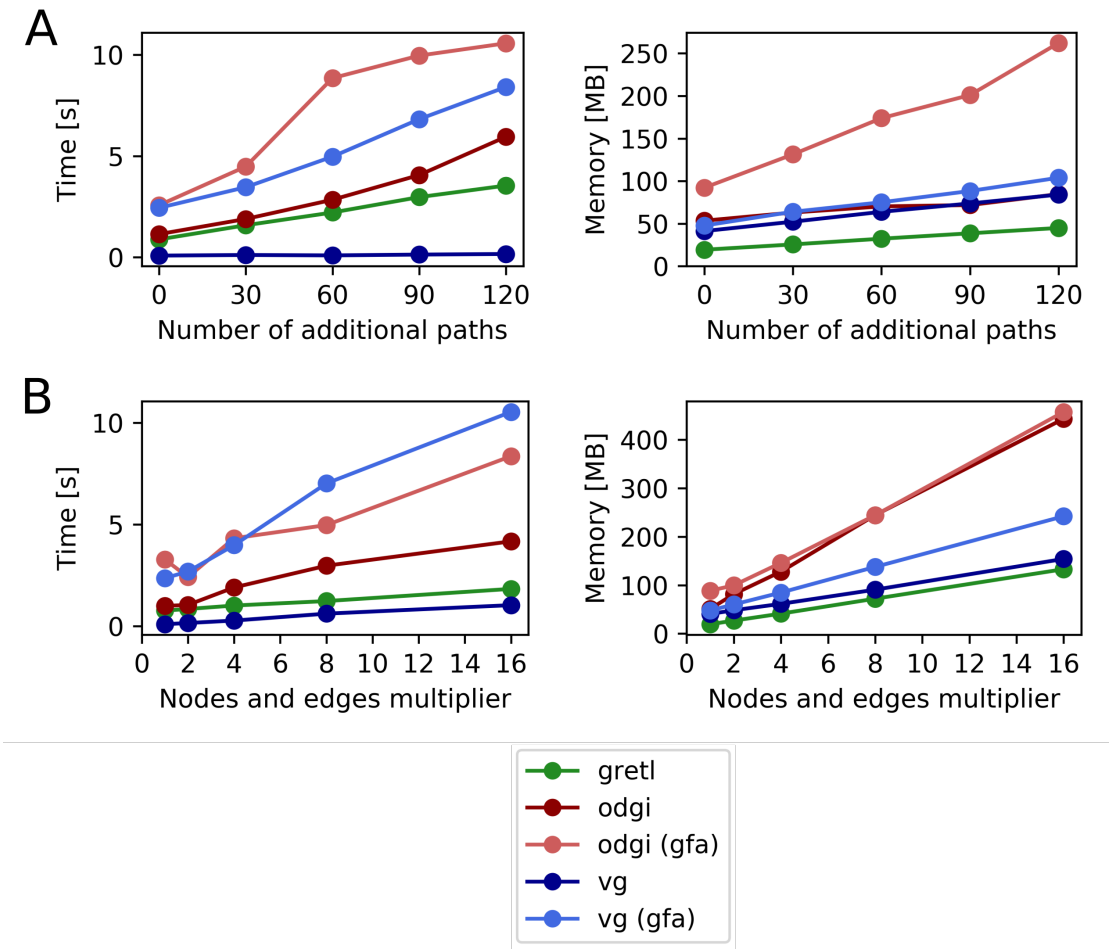


Fig. 2.2 **Scaling properties.**
(A) Based on increasing number of paths. **(B)** Based on increasing number of nodes/edges. The analyses are based on chromosome V of yeast with 30 paths. The number of paths in (A) was increased in every step by adding all paths from the base graph. For B, we repeatedly (indicated by the multiplier) added the nodes and edges of the base graph to the evaluated graphs. The constant time complexity (A, left) of the vg tool (on native format) is likely due to not processing path-related information.

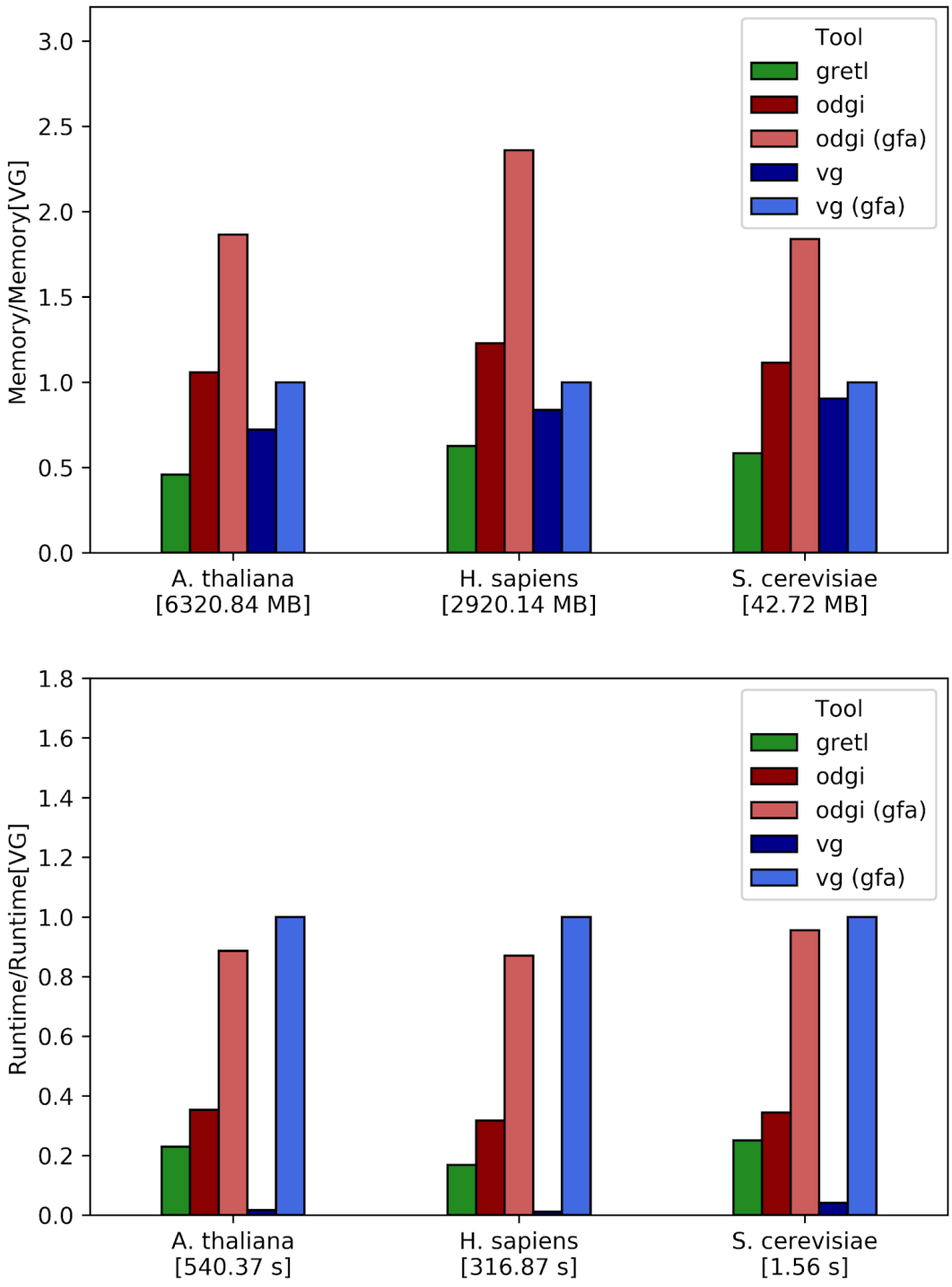


Fig. 2.3 **Run-time and memory benchmarking for `gretl`, vg stats and odgi stats in relation to vg stats.**
We added benchmarks for odgi and vg using their exclusive data formats as inputs. vg's (gfa) runtime and memory consumption are indicated in brackets at the bottom.

2.3.2 Comparing different parameters

Initially, `gretl` was developed for graph comparison within a grid search of parameter settings to identify the 'best' configuration.

When using the same base/input sequence (e.g. a set of genomes), it becomes straightforward to analyze the resulting graphs, as the genomic background of the input sequence can be entirely ignored. In our experiment, we focused on 30x *Saccharomyces cerevisiae* datasets, which represent 'simple' genomic backgrounds with manageable genome sizes. We conducted a representative experiment using the `PGGB` workflow, in which we adjusted the most critical parameters within the pipeline (Table 2.3). In total, we generated graphs by modifying five different parameters, resulting in 108 possible combinations.

Since we extract multiple statistics from one graph, a good visualization is a heat-map which can represent all computed features and all graphs at the same time. However, since the color grading value range in heatmaps is consistent for all displayed values, it is necessary to normalize the dataset. Hierarchical clustering was employed to identify graphs with similar statistical profiles. We recommend performing scaling by the maximum value per feature, resulting in values ranging from 0 to 1 (Figure 2.4). To represent the different graphs, an additional heatmap was included below, displaying the various parameter settings. Three colors were selected to indicate high (red), medium (blue), and low (green) parameter values (Table 2.3).

The most prominent parameter exhibiting a clear pattern is 's', which adjusts the segment length during the initial *wfmash* alignment step. Based on the clustering, which treats all statistics equally, features measuring "inverted nodes/edges" in any form are particularly influenced by the parameter 's'. Notably, the "inverted" feature shows a strong negative correlation with 's', with a Pearson correlation coefficient close to -0.8 (Figure 2.5). Positive correlations were generally weaker. However, a clear trend was observed whereby higher 's' values corresponded to higher overall similarity.

Another particularly relevant parameter is 'p', which controls the *mash* threshold between two segments in *wfmash* (Table 2.3). While detectable in the heatmap, the effects of 'p' are even more apparent through direct correlation analysis. A very high correlation was found for "Path node size" in base pairs, affecting both the standard deviation and average values. General graph statistics are also influenced: "Node size" exhibits the highest correlation with 'p', along with overall graph size in base pairs. This is consistent with theoretical expectations: a higher 'p' threshold reduces the number of alignments, resulting in a less condensed graph with larger nodes. Negatively correlated features include those related to "compression" and node count. As nodes grow larger, fewer are needed to represent the

graph, and thus compression decreases; the input sequence remains constant while the overall graph size expands.

The parameters 'n' (number of secondary alignments) and 'k' (seqwish initial seed hits) were also found to influence graph structure. Although these were not investigated in detail here, distinct patterns can be observed in the heatmap. The parameter 'asm' (alignment settings) appeared to have no substantial effect on graph structure.

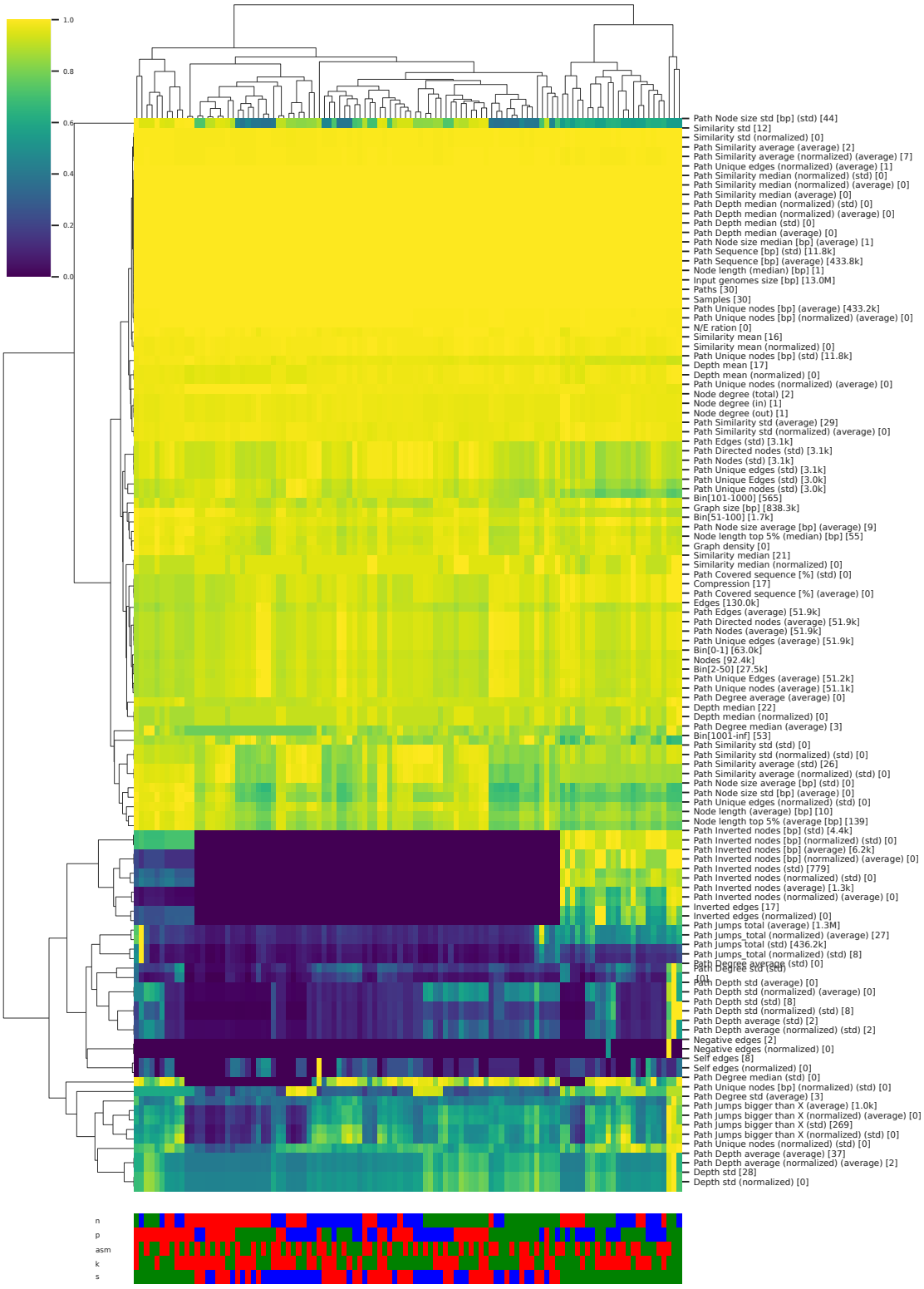


Fig. 2.4 **Graph statistics across yeast graphs.**
This heatmap visualizes graph statistics for each yeast graph constructed using different parameter sets. Rows and columns are hierarchically clustered. Colors at the bottom of the plot show different parameter settings, each row represents one parameter. High values of the parameter are in red, middle in blue and low in green.

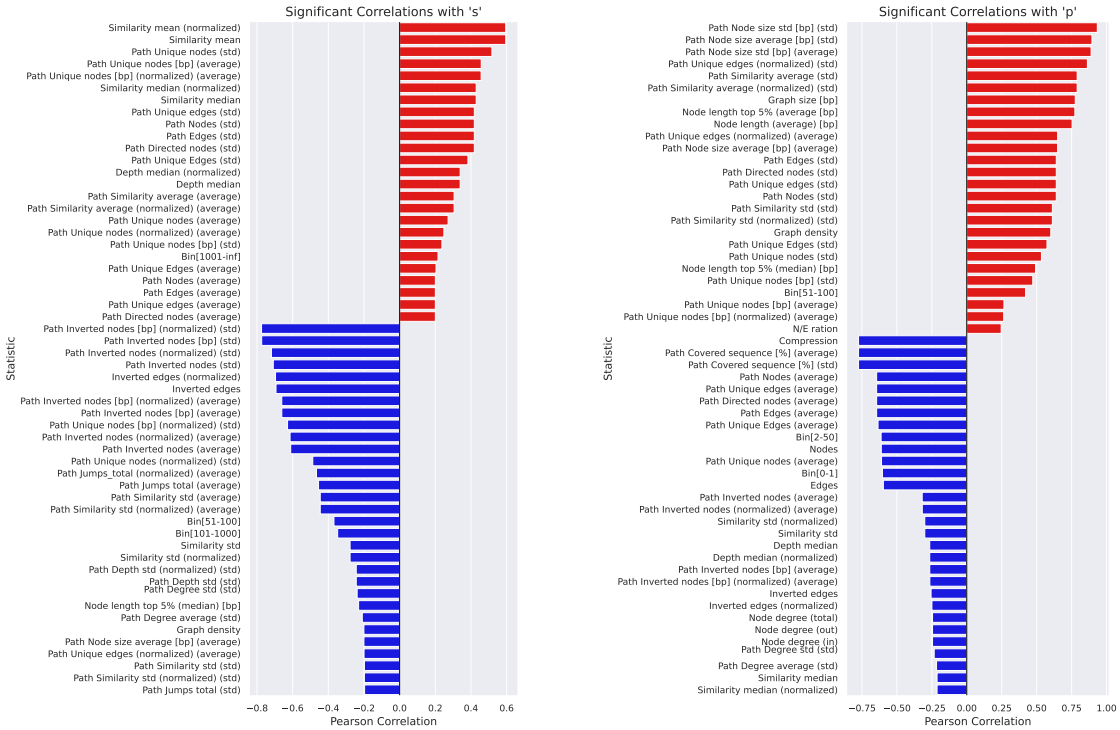


Fig. 2.5 **Correlation between** PGGB **parameters and graph statistics.**
Features significantly correlating with 's' **(A)** and 'p' **(B)**. Red bars indicate positive, blue bars negative correlations, with a vertical line at zero for reference. Only those features significantly affected by the parameter ($P < 0.05$) are presented.

For a more detailed perspective, a scatter plot (2D) representation can be employed (Figure 2.6). This approach facilitates the identification of the effects of parameter choices and enables the examination of parameter-specific statistical patterns more effectively. Using different colors or markers the user can highlight certain parameters for better identification of the right one. In this case, we used node degree and average node length (in base pairs), with the 'n' (number of secondary alignments) and 'p' (mash threshold) parameters represented by color and marker symbol, respectively. Even by eye, we can observe that graphs constructed with a strict/high 'p' parameter generally exhibit a low node degree and a high average node length. The number of secondary alignments ('n') typically has little effect on the overall structure of the graphs, as such alignments are present throughout. However, lower values of 'n' appear to result in graphs with a generally low node degree.

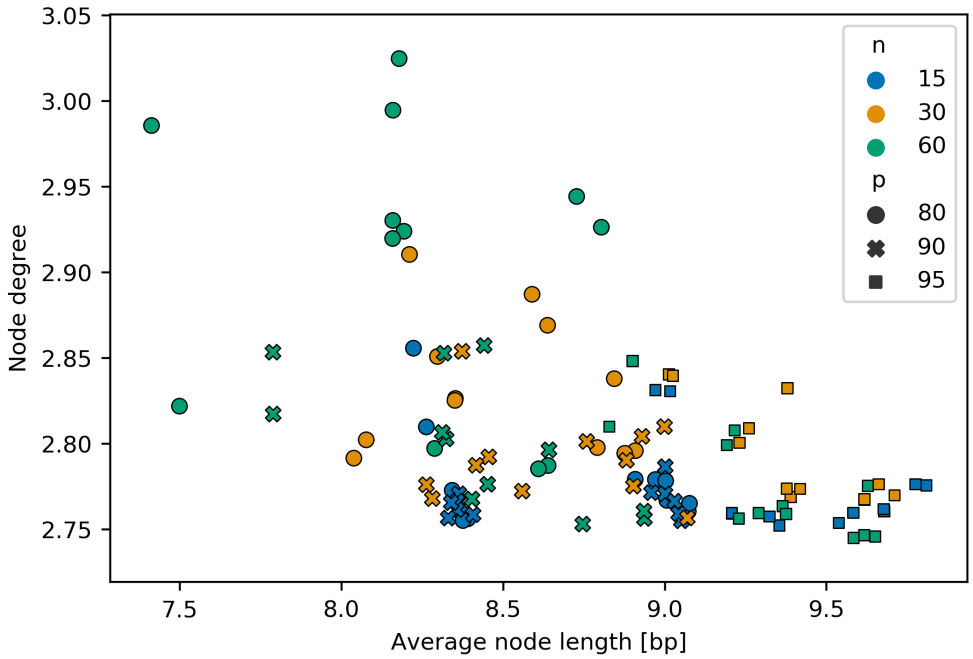


Fig. 2.6 **Relationship between node degree and the average node length in base pairs for all graphs built with different combinations of parameters.**
The different colors highlight the "percent identity" in the wfmash step ('p'), the different shapes the (secondary) n-mappings ('n') of PGGB. Graphs are based on chromosome IX from 30 *S. cerevisiae* genomes. Graphs built with different 'p' can easily be distinguished and this parameter seems to have a strong influence on the graph.

2.3.3 Comparing different organisms

Using `gretl` it is also possible to compare variation graphs from different species or populations. Unlike comparisons within the same dataset, such as parameter optimization or method selection, these graphs originate from distinct genomic backgrounds, and these differences should be reflected in their graph properties. Whether the goal is to identify these distinctions at the graph level or create uniform graph structures for comparing basic statistics, this is an integral part of the parameter selection process.

Nevertheless, the genomic architecture of a species is influenced by a multitude of factors. These factors can include genetic drift, mutations, genetic bottlenecks, or natural selection, all of which impact the genetic architecture of the species. Furthermore, different chromosomes within a species may have undergone distinct evolutionary processes and could originate from different ancestries.

Differences extend not only to the complexity of the genomes themselves but also to variations between individual populations. For instance, many bacterial genomes (strains) exhibit high collinearity and minimal differences between them. In eukaryotes, particularly in plants, transposable elements and other genomic processes can lead to intricate rearrangements, duplications of transposable elements, and gene capture events. These complexities pose new challenges in terms of comparison and alignment. When comparing graphs from different species using `gretl`, it is recommended to exclude all non-normalized statistics to reduce inflation or false representation of features. Nevertheless it is helpful to have a look on non-normalized features, if genomic input is very similar.

As previously mentioned, this experiment employed graphs from *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, and *Homo sapiens* to facilitate a species-level comparison at the graph level. Each species is represented by five chromosomes, each with varying quantities of samples and paths (Table 2.1).

To emphasize the comparison rather than the absolute (normalized) values, we standardized all measurements as ratios relative to the highest observed value for each specific statistic. This results in a shifted representation where individual values are portrayed as one (100%), even if their original absolute values are substantially lower. The clades within the resulting dendrogram accurately reflect species associations. When selecting two specific statistics, such as average node length and node degree, and visualizing them in a two-dimensional scatterplot, distinct species-specific patterns become apparent, highlighting structural differences in the underlying genomes (Figure 2.7).

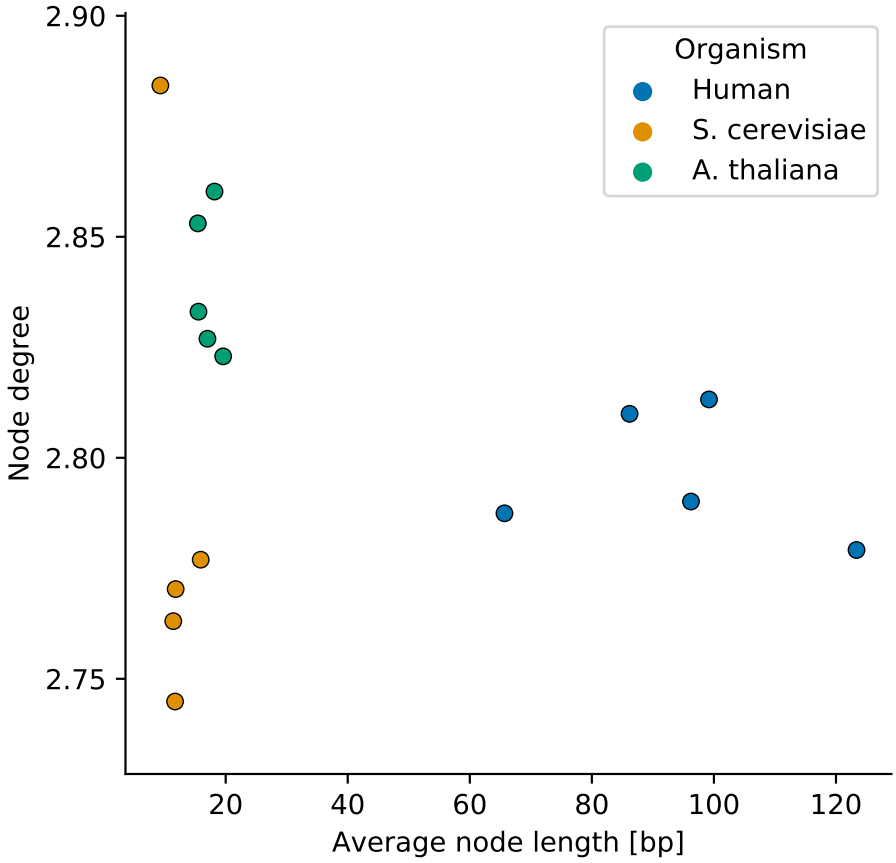


Fig. 2.7 **Species-wide comparison**
Scatter plot showing two selected statistics (node degree and average node length [bp] across multiple graphs, enabling comparisons between different species. Distinct patterns can be observed not only between species but also among individual chromosomes, highlighting species- and chromosome-specific graph characteristics.

2.3.4 Minigraph – PGGB

In this study, a comparative analysis was conducted on genome graphs published as part of the Human Genome Draft Pangenome [53]. These graphs were generated using the same set of human samples but were constructed through distinct methodological approaches.

As outlined previously, the two graph construction methods are based on fundamentally different principles. Nonetheless, both yield fully resolved structures that represent the input genomic data.

Notably, the graphs tended to cluster according to the construction method rather than by chromosome, indicating that methodological differences are of greater influence on graph structure than chromosomal origin.

A prominent distinction between the two graph construction approaches is observed in metrics such as average node length, inverted edges, and overall graph compression. MC graphs generally exhibit lower average node lengths, particularly in the top 5% percentile, but in contrast, show a higher number of inverted edges and reduced compression. MC graphs also tend to display higher similarity values; however, this comes with increased standard deviation, indicating less consistency across the graph.

In contrast, PGGB graphs contain almost no inverted edges, making this a distinguishing feature of the MC method. One notable exception is PGGB's chromosome 22 graph, which differs substantially from all others by showing a high number of negative edges, extremely long node lengths, and the highest count of self-edges. This makes chromosome 22 a consistent outlier, particularly in the PGGB set. The behavior of chromosome 22 as an outlier suggests that certain genomic regions, possibly due to structural complexity or repetitive content, are more sensitive to graph construction parameters.

Despite these differences, both graph types maintain a similar ratio of nodes to edges and share an almost identical median node length, likely reflecting the prevalence of SNPs, the dominant form of variation in the dataset. While many statistics are broadly comparable, the mentioned differences underscore method-specific biases.

Interestingly, MC graphs appear more internally consistent across chromosomes, as reflected in their tighter clustering in the dendrogram. This consistency may make MC graphs more suitable for comparative studies where reproducibility across chromosomes is important. By contrast, PGGB graphs, largely due to the outlier behavior of chromosome 22, show greater variability, resulting in divergent subclusters even within the same construction method.

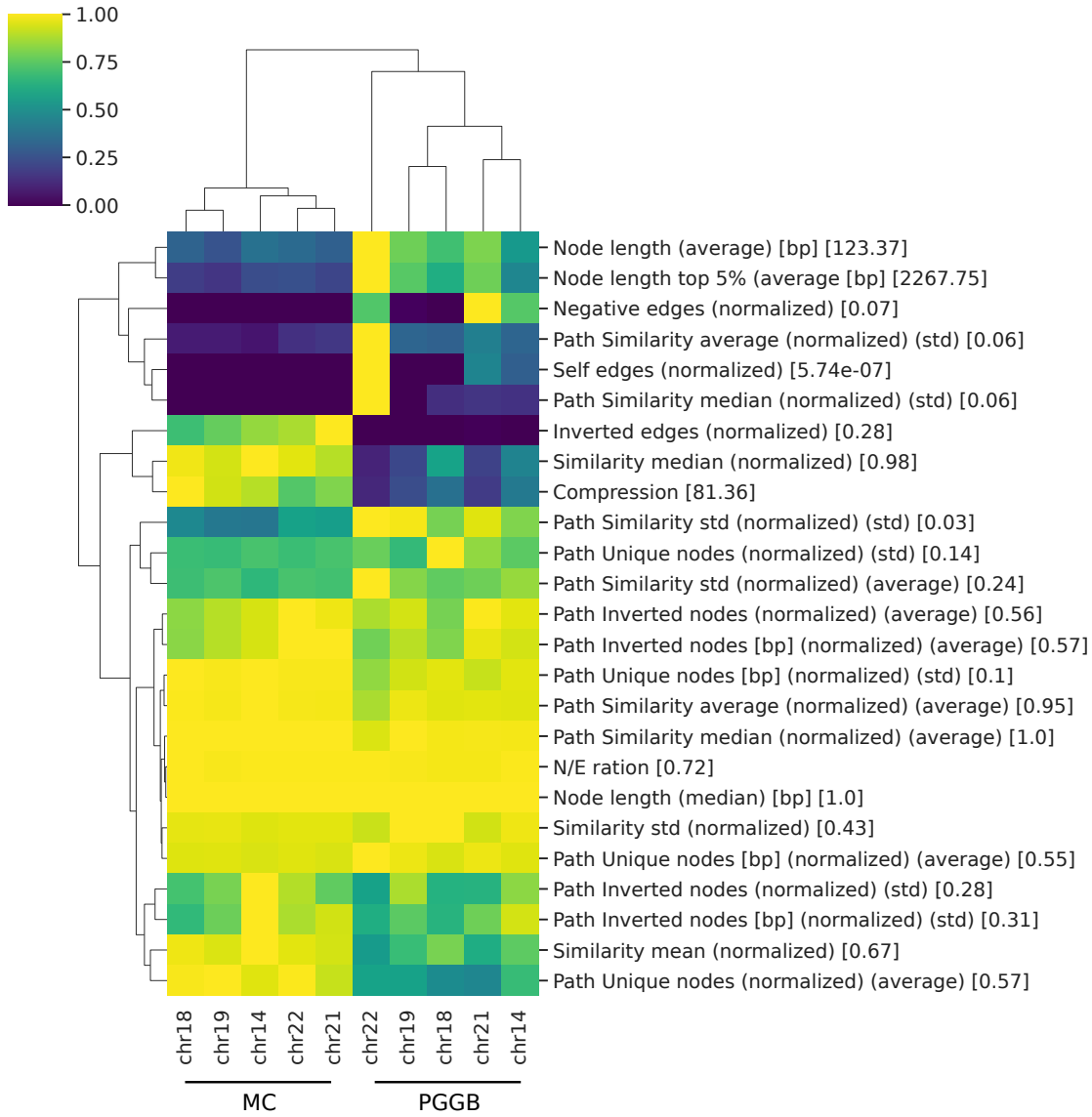


Fig. 2.8 `gretl` **stats - comparison of different methods.**
Individual graphs cluster by method and not by chromosome. Only normalized metrics are shown here. In addition, values were scaled by maximum for each feature (rows). Maximum values shown in brackets.

2.3.5 In-depth analysis

Analyzing a genome graph can be challenging, especially when the features of interest are not easily detectable. Most analyses provided by this tool focus on the pangenome, including its classification and localization. Additionally, the tool enables the identification of regions with varying levels of complexity within the graph, allowing users to pinpoint and link regions of interest. To demonstrate these capabilities, we performed several analyses on the

Saccharomyces cerevisiae (yeast) and the human genome, as shown below. A more detailed examination of our methods is presented in Chapter 3, which focuses on the analysis of variation graphs constructed from 28 genomes obtained from *Arabidopsis thaliana*. Selected, specific analyses are highlighted in this section.

Path-level comparison of human variation graph

In this study, we used `gretl` to analyze the `PGGB` graph of *H. sapiens* Chromosome 18, constructed from 48 individual samples annotated by their respective superpopulations. While each superpopulation exhibits statistically distinct graph profiles, the samples are broadly distributed, particularly those from the African (AFR) group. However, this analysis is intended as a representative example rather than a comprehensive population-level comparison. Due to the limited and uneven sample sizes across groups, no definitive biological conclusions should be drawn from this dataset alone.

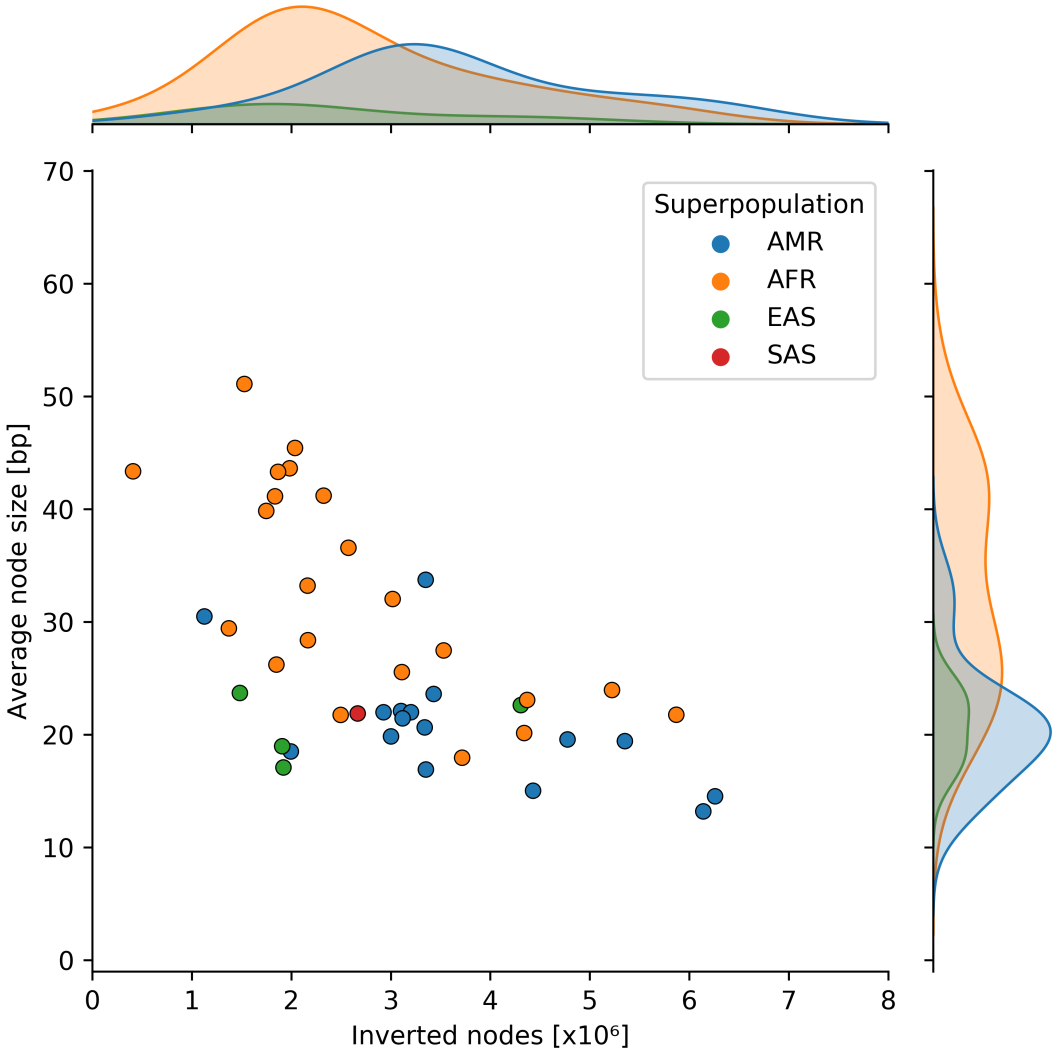


Fig. 2.9 **Relationship between average node size and the number of inverted nodes of each path in the H. sapiens chromosome 18 (hprc) graph.**
The path names are annotated to their superpopulation: AMR: Admixed American, AFR: African, EAS: East Asian, SAS: South Asian.

Complexity analysis in *Saccharomyces cerevisiae*

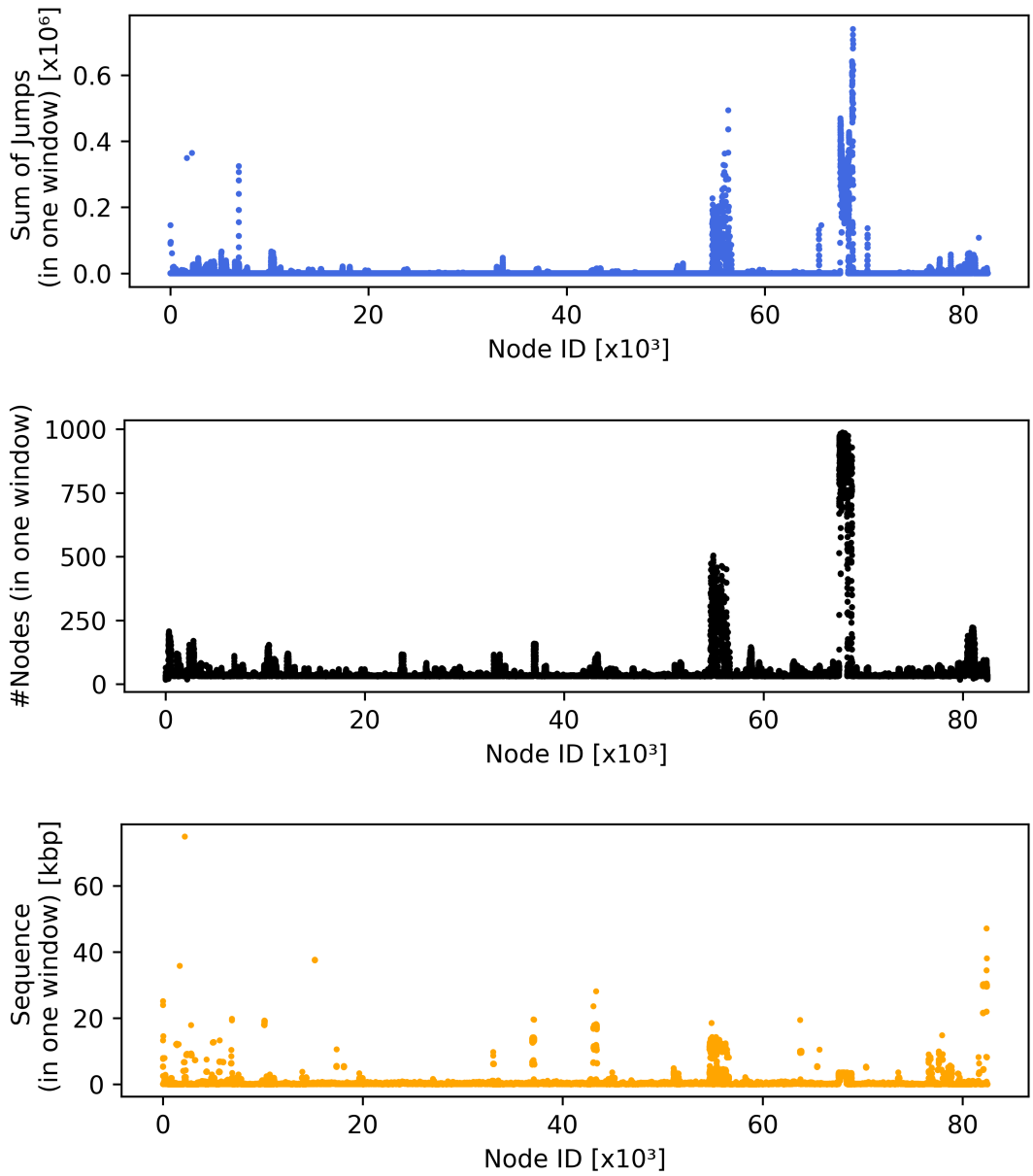


Fig. 2.10 `gretl` **nwindow - Detection of regions of high local variability.**
Graph-based window approach iterating over each node in the graph and capturing all nodes in up to 10 steps away. Each window is summarized by amount of sequence (top), number of nodes (middle), or summary of node ID distance (bottom) from the starting node.

Correlation

Our tool provides several statistical measures designed to identify nuances between different graphs. However, in many cases, these statistics reflect the same underlying feature in the variation graph. As such, they may be aggregated into a single value or, at the very least, interpreted as related metrics that describe the same aspect of the graph structure.

To do so, we examined how other parameters affect individual graph statistics and identified the major driving factors. We focused on the statistics themselves, specifically, which statistics correlate with each other. Using our previously described parameter dataset in yeast, we analyzed the correlations among these statistics, specifically looking for strong (anti-)correlations between different measures when comparing graphs from the same dataset.

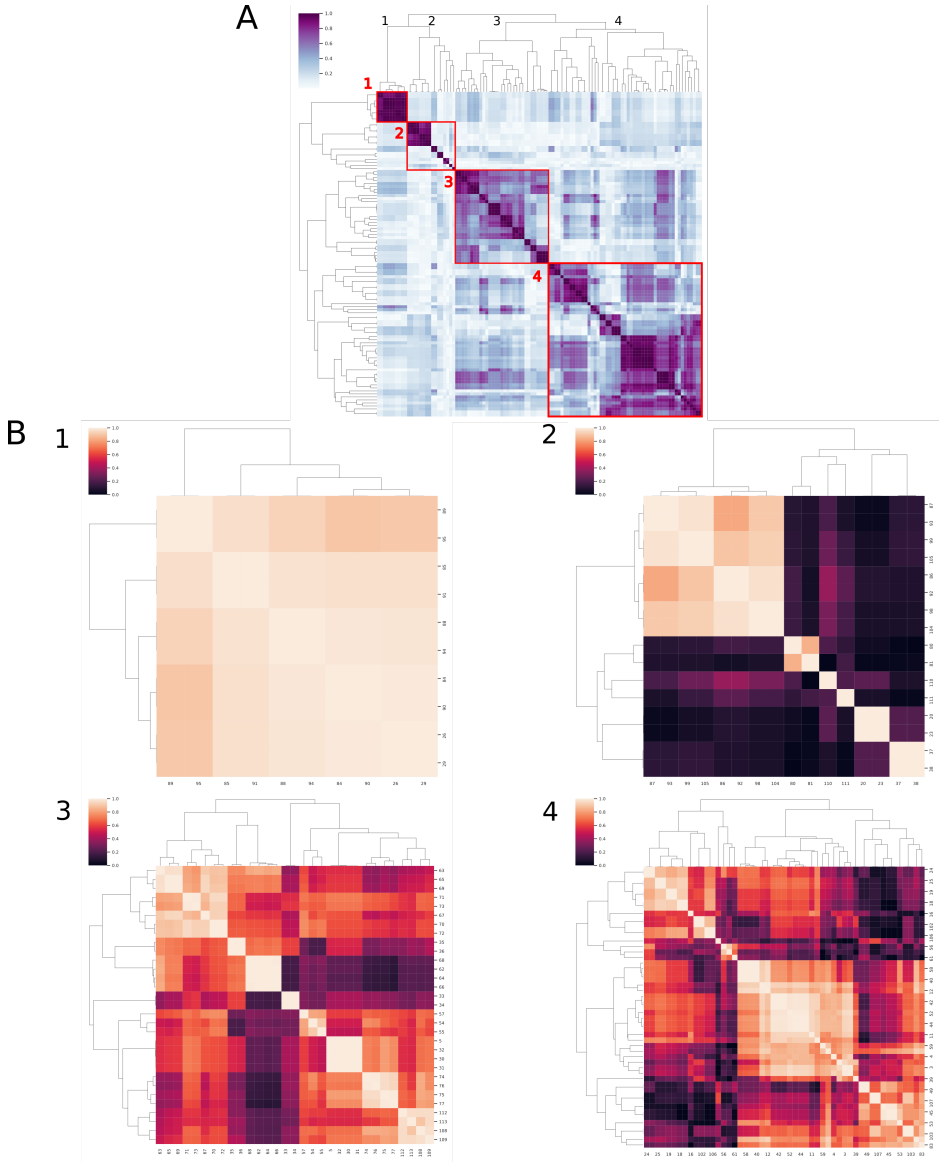


Fig. 2.11 **Correlation Analysis of Yeast and *Arabidopsis* graph statistics.**
This heatmap presents the correlation coefficients derived from various graph statistics, which were calculated through the analysis of a yeast variation graph. The visualization highlights the interrelationships among different metrics, revealing significant associations between them.

Hierarchical clustering based on correlation was used to identify clades of similar features. While multiple clade groupings could have been defined, determining an optimal threshold proved challenging. Ultimately, we selected a threshold that resulted in four distinct clades.

2.3.6 Workflow

We also provided an example on how to use a combination of our new graph analyses. This example had the goal to “Identify highly variable regions based on genome graphs” (Figure 2.12). The target is a specific statistical profile that limits possible candidates for highly variable regions to a reasonable number. After finding the optimal graph (yellow), we continue with several analysis steps (orange). Finally, the regions of interest can be compared to existing annotations, e.g., by BLAST, or the subgraph can be extracted and investigated in detail. A combination of both might be the optimal approach (green). Here, we identified a double loop structure caused by presence-absence polymorphism resulting in a highly variable region in close proximity to the bifurcation/merging point (Figure 2.12).

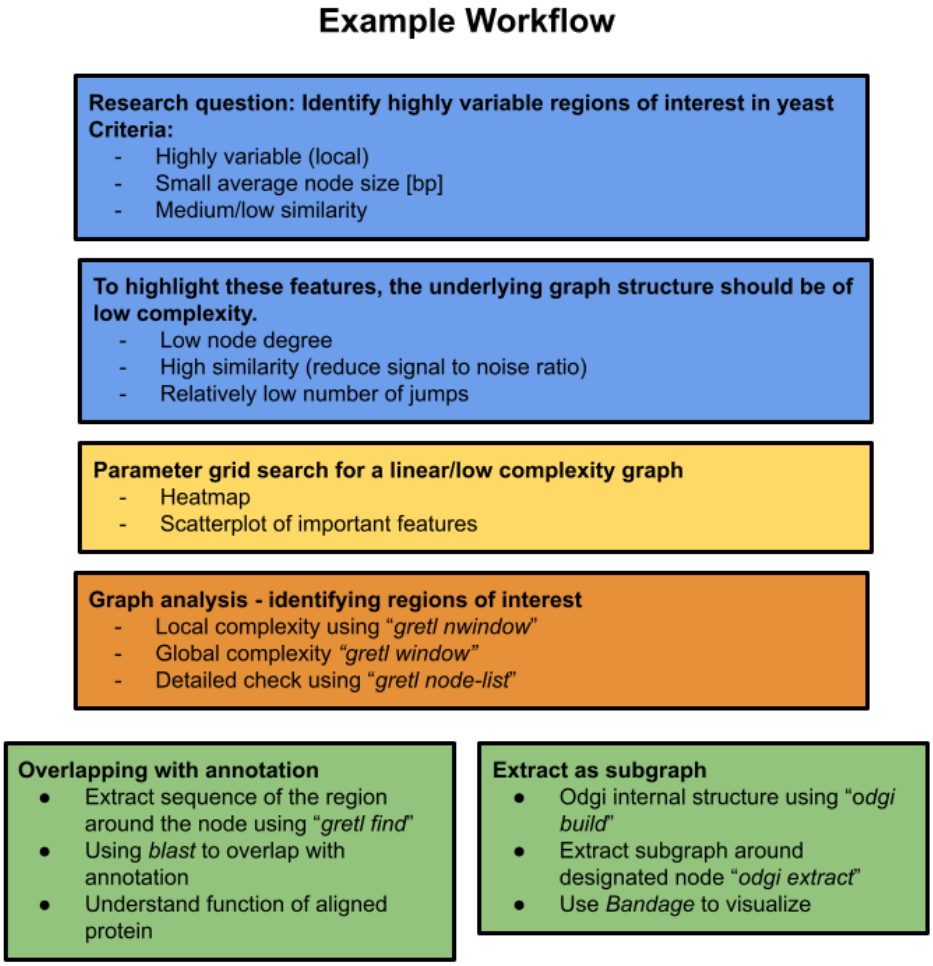


Fig. 2.12 **Example workflow: "Identify highly variable regions based on genome graphs"**

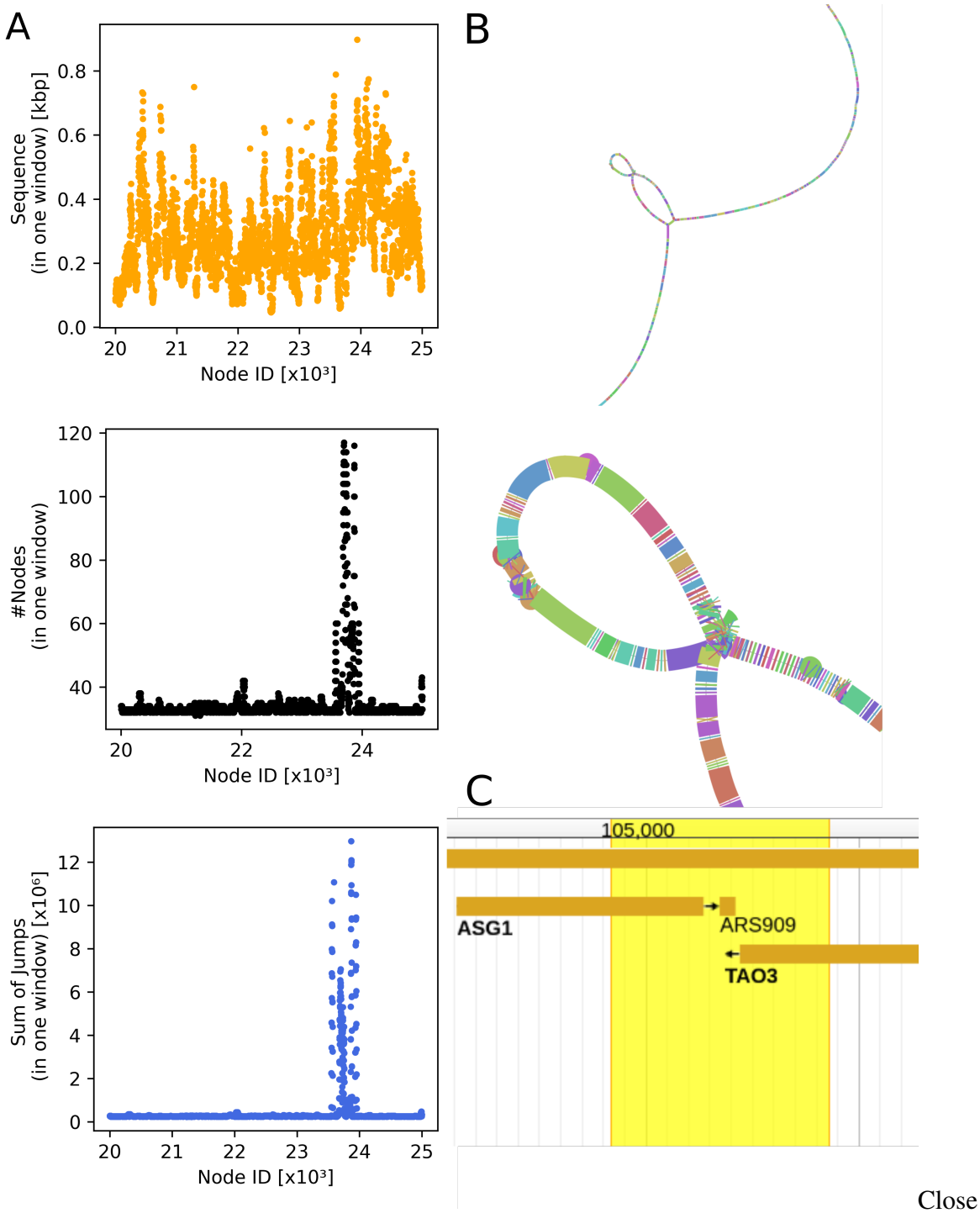


Fig. 2.13 **Example results.**
(A) Based on the nwindow analysis, nodes with an ID at 23850 showed a dense pattern with no significant increase in sequence. **(B)** The nodes were identified to be in the middle of a double loop structure close to a SNP array, which caused the high number of nodes. **(C)** We are not able to comment about the origin of the highly variable region, but from a sequence similarity search with the loop sequence to the reference annotation, we found an autonomously replicating sequence bordered by a regulatory and a signal transduction gene.

2.4 Discussion

Our parameter study demonstrated that evaluating graphs built with different parameter settings is crucial. Among the tested parameters, 's', which defines the segment length during the initial *wfmash* alignment step, was especially influential. This parameter plays a critical role because alignment forms the foundation for graph construction via *seqwish*. Longer segments reduce the likelihood of spurious or repetitive alignments, leading to graphs with fewer cycles and a more linear structure. In contrast, shorter segments increase the resolution and enable the alignment of repetitive or mobile elements such as transposable elements, motifs, or even genes. However, shorter segments also increase the chance of redundant or ambiguous alignments, contributing to graph complexity and collapsed structures.

The effect of 's' is strongly dependent on its interplay with parameter 'p', which sets the *mash* similarity threshold. Only segments surpassing this threshold are aligned. A combination of low 's' and low 'p' increases sensitivity but can lead to over-collapsed graphs, as shown in our results. In contrast, using high values for both parameters yields more linear, conservative graphs with fewer complex features. Interestingly, 'p' alone can have a similar simplifying effect on graph structure by limiting the number of accepted alignments, thereby increasing node size and reducing graph density and compression.

The parameter 'n', which defines the number of secondary alignments per segment, was expected to have a larger effect. While we did not observe a strong correlation, we hypothesize that 'n"s influence is reduced due to the all-vs-all alignment approach and the way *seqwish* merges overlapping alignments into the graph. Lower 'n' values may omit certain alignments, but in densely connected regions, indirect alignment links likely compensate. Higher 'n' values may only matter when a segment appears in multiple divergent regions, more common for short segments, again highlighting the indirect dependence on 's' and 'p'.

Parameters 'k' (the minimum seed size in seqwish) and asm (alignment parameter) had minimal impact. The former was expected to promote graph linearity by filtering out weak alignments, but this effect was not observed. The latter, surprisingly, showed no measurable influence in our current setting. While 'asm' controls high-level alignment behavior in the normalzing stage (smoothxg), it may only affect graph structure under specific conditions, such as complex genomes or poorly tuned alignment parameters. Previous versions of `PGGB` have shown that default 'asm' values can cause fragmentation, SNP arrays, or large unaligned regions, but such behavior was not evident in our yeast-based test case.

In summary, parameters 's' and 'p' exert the strongest and most consistent influence on graph structure. Their effects are both independent and combinatorial, shaping key features such as size and graph linearity. The remaining parameters contribute marginally or under

more specific conditions, reinforcing the idea that careful assessment of alignment-related settings is essential for producing interpretable and stable graph topologies. We expect that in more complex genomes, the effects of parameter variation would be even more pronounced than those observed in our experiments. Although the full parameter space and their combinations are not yet systematically defined, the parameters discussed here are considered among the most influential by developers and users alike.

Complementing our parameter study, we also observed substantial differences in graph construction between Minigraph-Cactus (MC) and PGGB. Notably, even when using identical input sequences, the resulting graph structures varied considerably depending on the construction tool. These tool-specific differences far exceeded the effects of parameter variation alone, suggesting that aligning graph outputs across tools may not be feasible, even with extensive parameter tuning.

A key reason for these discrepancies is the way MC handles sequence divergence. Unlike PGGB, MC does not construct graphs at base-pair resolution. In regions of high divergence, the MC algorithm tends to omit non-reference sequences, retaining only the reference backbone. In some cases, entire divergent segments are removed or samples are split, aligning only the most similar regions. This results in a more linear graph structure but at the cost of losing detailed variation in complex or repetitive regions.

This discrepancy may also reflect the inherently more linear architecture of MC graphs, potentially coupled with a greater abundance of repeat elements in PGGB graphs that reduce node similarity. Indeed, similarity was generally higher in MC graphs, though accompanied by greater variability. These structural differences may have practical consequences for downstream applications, including read alignment, variant calling, and functional annotation, where graph topology directly influences mapping accuracy and interpretability.

Interestingly, PGGB graphs contained almost no inverted edges, highlighting the distinct design principles behind each approach. While MC was originally optimized for human genome graphs, its default settings may not generalize well to more complex plant genomes. Nevertheless, in gene-rich regions with low transposable element activity, MC and PGGB graphs are expected to exhibit more similar structures.

These findings underscore the variability inherent in genome graph representations and highlight the critical importance of method selection in producing accurate and interpretable genomic models. The choice of graph construction method not only affects the structural properties of the graph itself but may also determine its suitability for specific analytical tasks. As graph-based genomics continues to expand, careful consideration of tool design and domain specificity will be essential for robust and reproducible analyses.

We contribute `gretl`, a fast, efficient, and user-friendly stand-alone tool for generating a wide range of statistics and insights into the structure and composition of genome graphs, complemented with a set of user-friendly Python scripts for downstream analyses. `gretl` generates 108 different metrics for a single variation graph. We highlight path-centric statistics and analyses especially designed for genome graphs that have not yet been implemented by other tools. It is important to note that the quality of the genome assemblies used to generate the genome graph can significantly affect the accuracy and completeness of the generated metrics and subsequent downstream analyses. As such, it is essential to carefully evaluate and validate assembly quality before using `gretl`. In our experience, the building of graphs from complex genomes such as those of plants is highly affected by parameter choice. While `gretl` can process any graph which adheres to the GFAv1 specification, it is required that node IDs are numeric, and a sorted ID space is necessary for all "Jump"-related statistics. We recommend using path-guided 1D SGD ordering, which can be achieved effectively using the 'odgi sort -Y' functionality during the preprocessing stage [46, 51]. `gretl` is unique in that it provides both graph-based and path-based statistics, allowing users to gain insights into both the overall structure of the genome graph and the specific paths/samples through the graph that correspond to genetic variation. Finally, `gretl` is designed to be modular and extensible, allowing for the future addition of new features and statistics.

Chapter 3

Application of genome graphs – Pangenomic analysis of the 1001G+ dataset

Parts of this chapter have been published in *Towards an unbiased characterization of genetic polymorphism*, available on *bioRxiv* at https://doi.org/10.1101/2024.05.30.596703.

3.1 Introduction

In the upcoming chapter, several tools mentioned earlier have been utilized, with a particular focus on analyzing the 1001 Genomes Plus (1001G+) graph. The 1001G+ project is a successor of the original 1001 Genomes project [3], but is now including only a few accessions from each of the eight admixture groups. In addition, we added ecotypes from recently published papers, which included samples from China, Madeira, and Africa [27]. The genomes have been sequenced using Complete Long Reads (CLR) technology by the Max Plank Institute for Biology Tübingen and the Gregor Mendel Institute of Molecular Plant Biology Vienna. Scaffoldings of several accessions have been supported by manual curation and optical maps from Corteva resulting in chromosome-scale assemblies. Subsequently, we constructed genome graphs in a chromosome-wise way to reduce complexity and remove unnecessary linkage between chromosomes. Our genome graphs were constructed with PGGB, a robust and reference-free approach developed by Erik Garrison and Andrea Guarracino [39]. In total, we generated five graphs with 28 paths, 17.47 M nodes and 24.12 M edges (s. Table 3.1.)

Our goal was to perform a reference-free analysis of the selected samples and a comparison

of graph-based structural variation analysis to Pannagram, a newly developed SV detection method [57]. We took efforts to remove reference bias in the graph analysis, and perform reference-free scaffolding, annotation and expression analysis. However, in many aspects, the usage of a reference could not be avoided since many state-of-the-art workflows and methods require one. Nevertheless, our analysis aimed to show that reference-free approaches are on par with traditional workflows and that analysis on whole-genome level gives better insights into genomic regions which were not feasible before.

3.2 Materials and Methods

3.2.1 Materials

Summary statistics of the generated chromosome-level genome graphs are shown in Table 3.1. The ecotypes used to generated the graphs are described in Table 3.2.

Table 3.1 **Summary of the genome graphs.**
This table provides a summary statistics of the constructed graphs, including number of nodes and edges, graph size and compression. Each row corresponds to a distinct chromosome, while each column denotes a different attribute of the graph. Notably, all graphs feature 28 paths.

Chromosome	Nodes	Edges	Graph size [bp]	Compression
1	4,106,249	5,674,167	56,137,656	14.78
2	3,038,951	4,203,135	39,014,019	14.25
3	3,612,943	4,991,585	49,333,666	13.43
4	2,970,954	4,115,126	38,878,631	13.89
5	3,739,478	5,172,232	51,273,660	14.73

Table 3.2 **Table of the 1001G+ accessions.**
This table provides a summary of the ecotypes of the 1001G+ study, including identifier, sequence origin, country of origin, ADMIXTURE group and the indication of optical maps support. Each row is representing a different accession and each column corresponds to different feature of the accession.

Stock ID#	Identifier	Accession	Sequenced by	Seq. Tech.	PCR_free	Country	Admixture Group	Optical_map
CS78969	1741	KBS-Mac-74	MPI	CLR	yes	USA	germany	six_refs
CS76864	6024	Fly2-2	GMI	CLR	yes_GMI	SWE	south_sweden	
CS77137	6069	Nyl-7	GMI	CLR	yes_GMI	SWE	north_sweden	Corteva
CS77309	6124	T690	GMI	CLR	yes_GMI	SWE	south_sweden	
CS77384	6244	TRÄ 01	GMI	CLR	yes_GMI	SWE	north_sweden	
CS76778	6909	Col-0	MPI	CLR	yes	USA	germany	Kawakatsu_2016;six_refs
CS77266	6966	Sq-1	GMI	CLR	yes_GMI	UK	western_europe	Corteva
CS76941	8236	HSm	GMI	CLR	yes_GMI	CZE	central_europe	Corteva
CS77023	9075	Lerik1-4	GMI	CLR	yes_GMI	AZE	italy_balkan_caucasus	
CS76787	9537	IP-Cum-1	MPI	CLR	yes_MPI; YES	ESP	spain	Kawakatsu_2016
CS76886	9543	IP-Gra-0	GMI	CLR	yes_MPI	ESP	relict	
CS77133	9638	Noveg-3	GMI	CLR	yes_GMI	RUS	asia	Corteva
CS77279	9728	Stiav-1	GMI	CLR	yes_GMI	SVK	central_europe	
CS76581	9764	Qar-8a	MPI	CLR	yes	LBN	admixed	Corteva
CS77197	9888	IP-Pva-1	GMI	CLR	yes_GMI	ESP	spain	
CS78840	9905	Ven-0	GMI	CLR	yes_MPI	ESP	relict	
CS76366	9981	Angit-1	MPI	CLR	YES	ITA	italy_balkan_caucasus	
CS76405	10002	TueWal-2	MPI	CLR	YES	GER	western_europe	
CS929	10015	Sha	MPI	CLR	yes_MPI; YES	TJK	asia	
CS77369	10024	Tnz-1	MPI	CLR	YES	TZA	africa	Corteva
	22001	85-3	CAS	CLR	yes	CHN	china	Corteva
	22002	35-1	CAS	CLR	yes	CHN	china	
CS799913	22003	Taz-0	MPI	CLR	YES	MAR	africa	
CS799925	22004	Elh-2	MPI	CLR	YES	MAR	africa	
CS2107642	22005	Rabacal-1	MPI	CLR	YES	POR	madeira	
635AV	22006	Areeiro-1	MPI	CLR	YES	POR	madeira	Corteva
	22007	ET-86.4	MPI	CLR	YES	ETH	africa	

3.2.2 Methods

Modifying accession 22001

Initially, we discovered a very large reciprocal translocation in accession 22001 (alternative name 85-3) from the Yangtze River region, which swapped the distal portions of chromosomes 3 and 5 (Figure 3.1). We validated the translocation by PCR with two sets of primer pairs designed to either amplify the standard arrangement of chromosomes 3 and 5 of Col-0, or the two translocation junction regions in accession 22001. This rearrangement, which would presumably lead to decreased fertility in heterozygotes, appears to be quite rare as we did not identify other examples in a sample of 117 accessions sequenced with Illumina short reads from the same region [151]. For the purposes of this study, we manually rearranged this genome to match the ancestral organization. To identify the exact breakpoints, we aligned chromosome 3 of 22001 to all other sequences of chromosome 3 with minimap2 (`-x asm5`). After filtering the alignments to retain those longer than 50 kb (`fpa drop -l 50000`), we removed the sequence from the start to the first position of alignment and added the reverse complement to the end of chromosome 5. A collection of the scripts can be found at GitHub[1].

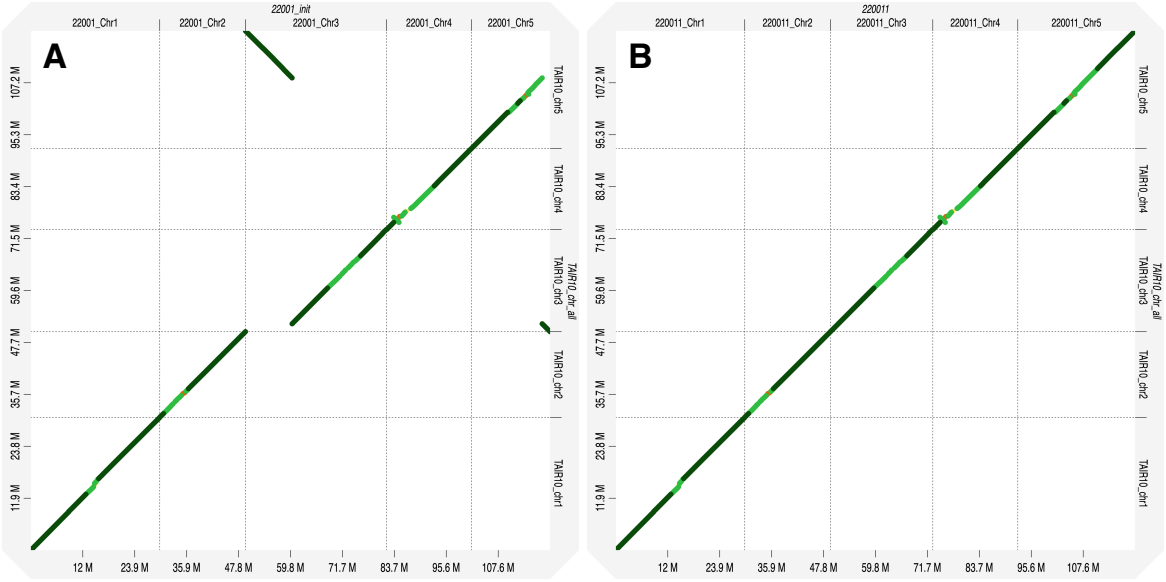


Fig. 3.1 **Reciprocal translocation in accession 22001.**
Dot plot of the original assembly 22001 (**A**) and of the modified assembly 22001m against (**B**) accession 22002. The translocation is readily seen at the beginning of chromosome 3 and the end of 5. Dot plots were created with D-GENIES.

[1]https://github.com/Gregor-Mendel-Institute/1001Gplus_paper/tree/main/02_analysis/11_mod_genome

Divergence Estimation

We utilized `mash` [104] to estimate the typical level of divergence among the accessions in our dataset. For this purpose, the `mash triangle` command was used to generate a lower triangular distance matrix. This matrix was subsequently visualized in Python using the `matplotlib` library. Due to our prior experience, where we observed that Mash distance tended to underestimate the actual divergence, we decided to conduct additional experiments involving SyRI and DeepVariant [43, 113]. SyRI was utilized for general structural variation (SV) detection among the various accessions, while DeepVariant was employed for SNP-based estimations. Specifically, SyRI was executed using the minimap2 alignment algorithm with the default parameters, as specified at https://schneebergerlab.github.io/syri/pipeline.html. DeepVariant was run with the default parameters, utilizing the `bwa mem` input alignment [76]. Principle component analysis (PCA) was performed with presence-absence data of our merged genome graph using `gfa2bin graph` on nodes. The PCA was made with the Python package `SciPy` and later plotted with `matplotlib`.

Graph statistics

All graph-based statistics presented in this chapter were computed using the `gretl` tool described in chapter 2. This framework enabled us to perform robust and consistent analysis across the five chromosome graphs, providing both general graph statistics and detailed pangenomic insights. Specifically, we used the following modules: `gretl stats` for basic graph-level statistics, `gretl ps` for path-related pangenome features, `gretl core` for core and accessory genome analysis, `gretl window` for local pangenomic complexity profiling, and `gretl bootstrap` to assess pangenome saturation. Saturation analysis with annotation was conducted using a custom Python script, which linked each graph node to one or more genomic features, including exons, genes, mRNAs, and intergenic regions. All visualizations were created using the `matplotlib` and `seaborn` libraries in Python 3.7.

3.2.3 Variation detection

Deconstructing the graph

Since the Bifurcation-Variation-Detection tool (BVD, see Outlook) was still under development as this analysis was done, the variation data was based on `vg deconstruct`. To achieve full perspective and cover all bubbles in the graph, `vg deconstruct` was run multiple times with each accession reference path once (`vg deconstruct -a -e`). Subsequently, the reported VCF files were converted to a BED file with information relevant to the anal-

yses provided. In addition, each chromosome was merged and genotype information was concluded and added. Bubbles were identified by start and end position, also reporting all traversals within these bubbles. To make it comparable to the provided data from Pannagram, structural variation bubbles have been categorized into single and multiple events (see below).

Bubble statistics

As described above, statistics for each bubble were collected based on their unique start and end positions. The resulting dataset included all traversals, represented by their respective node and edge sequences, as well as the number of distinct paths passing through each bubble. In an additional analysis step, we computed the size of each traversal by summing the lengths of the nodes it comprised, using their node IDs and associated sequence lengths. From this data, we extracted key metrics including the number of traversals per bubble, the number of paths, the maximum bubble size (i.e., the longest allele), and the bubble size ratio between alleles.

Pannagram

Pannagram is a whole-genome multiple-alignment pipeline[2] that produces an intuitive representation for genome-browser visualization [57]. This approach can be considered an extension of pairwise-alignment methods [43] to handle multiple genomes in a reference-free manner. We derived a pangenome coordinate system based on the resulting alignment and anchored it to the TAIR10 reference genome. It was run by our collaboration partner in Vienna.

The comparison between the two approaches was conducted using all structural variant (SV) data. Both datasets were filtered to retain only variants with at least one allele of size bigger than 14 bp. Additionally, SVs larger than 100 kb were removed to avoid technical artifacts and because such large variants could not be reliably post-processed. This filtering step primarily affected SVs detected in the graph-based dataset.

The filtered datasets, formatted as BED files, were then compared using `bedtools multiIntersectBed`. The resulting overlap file was analyzed using custom Python scripts. All scripts, input files, and intermediate results are available in the associated GitHub repository[3].

[2]https://github.com/iganna/pannagram

[3]https://github.com/Gregor-Mendel-Institute/1001Gplus_paper/tree/main/02_analysis/12_graph

Definition - Single and multiple events structural variation

In order to identify insertion-deletion (InDel) events, the detected structural variations were classified into two categories. The first category, Single Event Variation (SEV), reflects simple InDels and consists of one large and one small allele. The small allele is limited to a maximum size of 5 % of the size of the large allele, and the large allele can contain multiple traversals as long as they are highly similar to each other (at least 95 %). The second category, Multiple Event Variation (MEV), covers all structural variation bubbles that do not fit the criteria for a single event. Structural variation bubbles are defined as bubbles with a traversal size of at least 15 base pairs, regardless of the other traversal sizes. It is important to note that this definition only applies to SV bubbles.

3.3 Results

3.3.1 Divergence estimate

Within our dataset, Mash distance analysis revealed two distinct clades, as can be observed in Figure 3.2A. One clade consists primarily of samples from Europe and Asia, forming the core of the dataset derived from the original 1001G study. Within this clade, accessions of Asian origin, including several Chinese samples form a small, well-defined subcluster located in the upper left of the distance matrix.

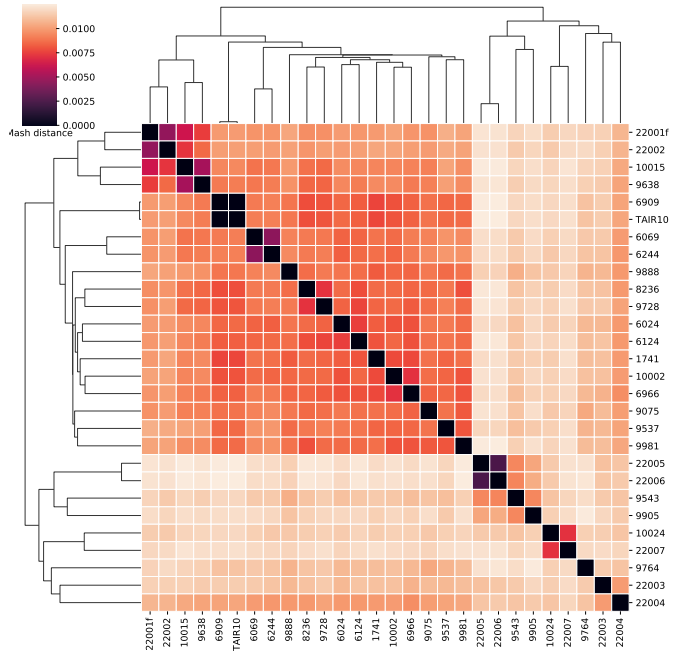
Notably, the two assemblies of the Col-0 accession, TAIR10 and our in-house resequenced version (6909), show an exceptionally small Mash distance, confirming their near-identical sequence composition. In addition, several closely related accession pairs were identified, suggesting highly similar genetic content. For instance, accessions 6069 and 6244, as well as 22005 and 22006, display strong similarity, likely attributable to their geographic proximity.

The second major clade predominantly consists of African, Madeiran, and so-called "relict" accessions. Within this group, clear subclades can be identified: one containing Madeiran and relict samples such as 22005, 22006, 9543, and 9905, and others comprising closely related African sample pairs such as (10024, 22007) and (22003, 22004).

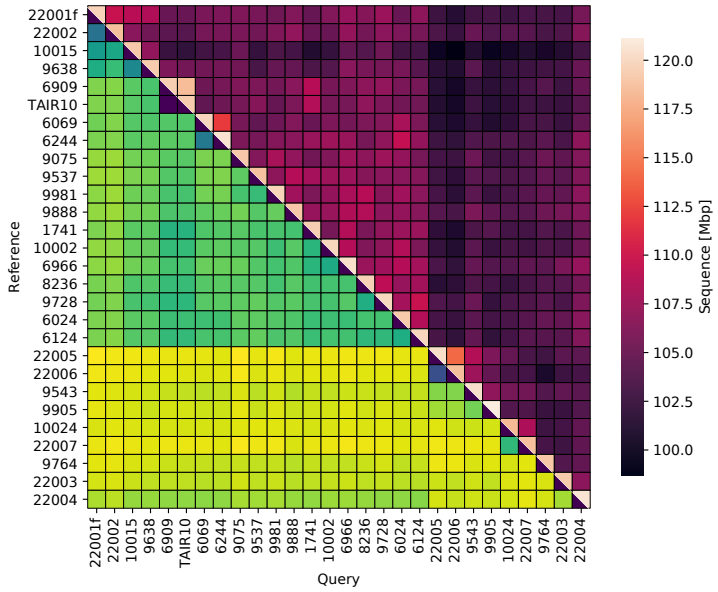
Since Mash distance can sometimes underestimate divergence in complex genomes, we also conducted experiments using SyRI and DeepVariant to evaluate syntenic sequence coverage and SNP counts, respectively. These two measures assess different aspects of genome comparison, thereby highlighting distinct underlying mechanisms. As shown in Figure 3.2B, similar patterns and clusters can be observed across all methods, supporting the consistency of our findings.

Another approach to infer common ancestry or to cluster samples is Principal Component Analysis (PCA). PCA is a dimensionality reduction technique that identifies key components explaining the greatest variance in the dataset, allowing samples to be projected into a lower-dimensional space while preserving important structural information. In our analysis, we observe a similar clustering pattern to what was seen in previous analyses (Figure 3.3). For example, TAIR10 and accession 6909 are positioned in close proximity.

The clusters described earlier are also evident in the PCA plot, which displays the first and second principal components. Most samples are distributed along the first principal component (PCA1), while only a few, namely accessions 22001, 22002, 10015, and 6938, show clear separation along the second component (PCA2). Interestingly, genomes not originating from Central Europe are generally more dispersed in the PCA space, whereas Central European accessions tend to cluster more tightly, suggesting shared ancestry or reduced genomic diversity within that geographic region.



(A) Mash distances. Lighter color represents higher distance and less similarity between the two genomes. Two clades can easily be found, representing mostly, a big group of European (non-relict) genomes and a outlier group of genomes containing accession of Madeiran, African or relict origin.



(B) SyRI synthenic regions (top/right half) and number of SNPs from DeepVariant (bottom-left half).

Fig. 3.2 **Divergence estimates of the 1001G+ genomes.**

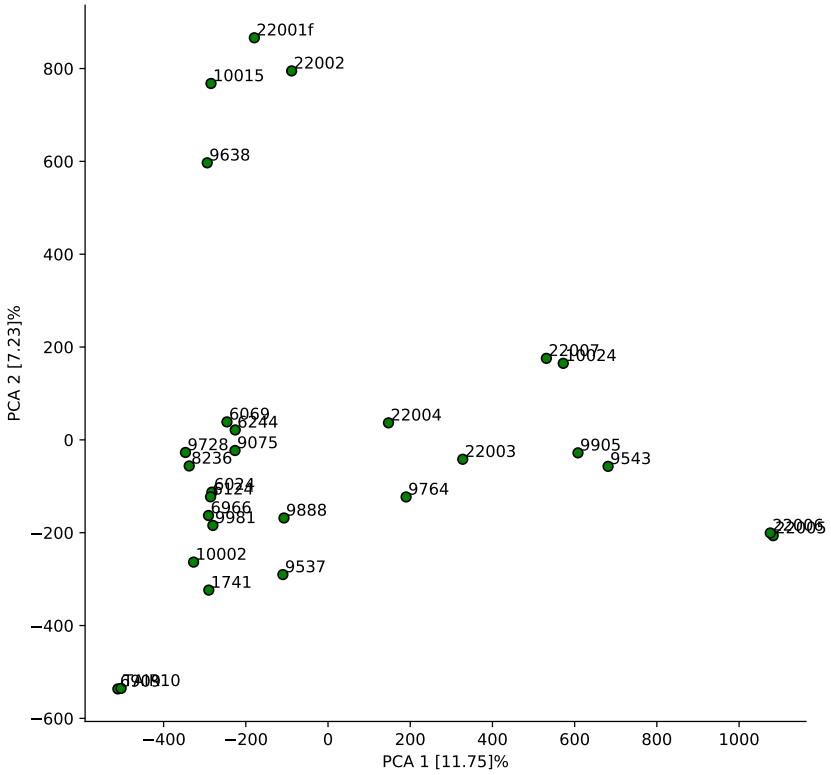


Fig. 3.3 **PCA based on node presence-absence.**

3.3.2 Graph statistics

Pangenome classification

The level of similarity between genomes offers an initial measure of genomic complexity in *Arabidopsis thaliana*. In our pangenome graph, total sequence content consists of 31.8% core, 25.4% private, and 41.7% soft-core regions. However, when considering node count, the distribution shifts: 67.2% of nodes are classified as soft, while only 17.1% and 15.2% are core and private, respectively. Despite the large total volume of private sequence, it accounts for only 1.8% of each accession on average, since private regions are specific to individual genomes and not shared. Conversely, core sequence, while comprising a smaller portion of the graph overall, represents 64.3% of each individual genome due to its presence across all accessions. These patterns underscore the asymmetry between collective genome content and its per-accession representation, reflecting how shared and unique regions contribute differently to graph complexity and individual genome structure.

When comparing the different accessions, the overall genomic variation appears relatively low. As shown in Figure 3.4, most genomes are not only similar in size but also contain

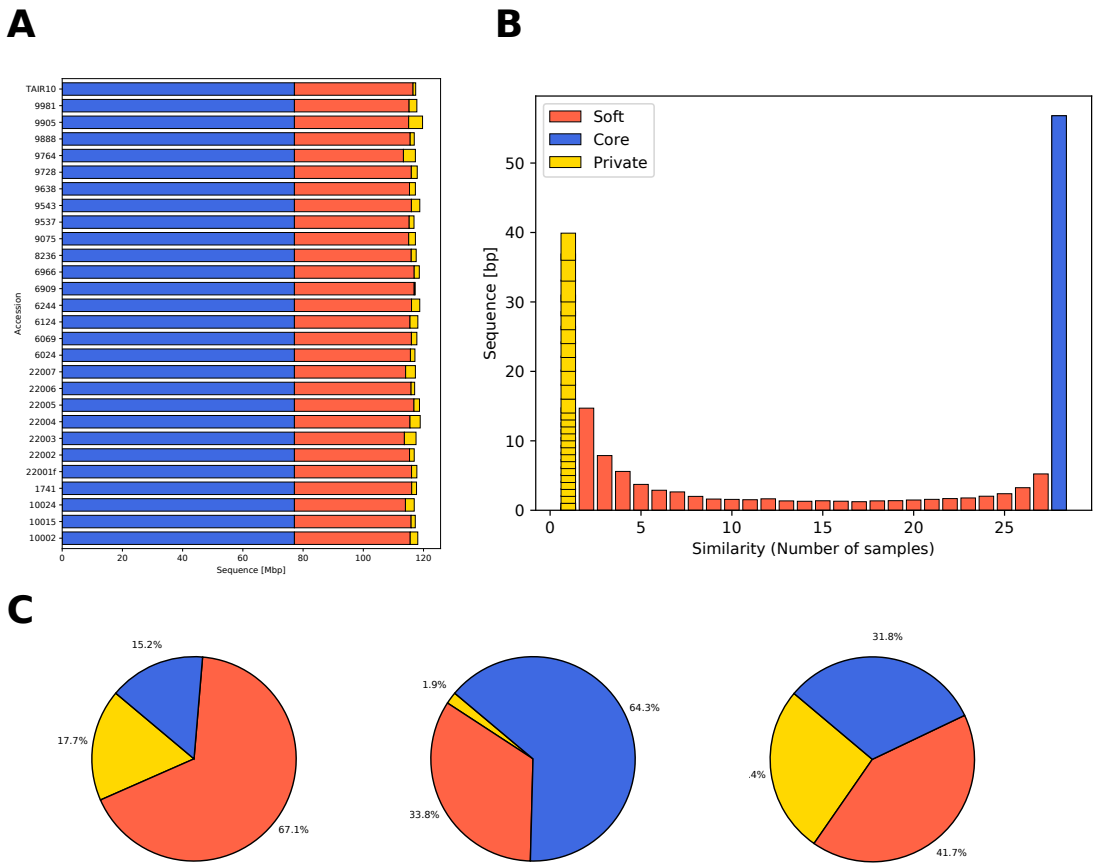


Fig. 3.4 **Composition of the pangenome in *Arabidopsis thaliana*.**
(A) The amount of core, soft and private sequence for each accession. **(B)** Composition of the pangenome visualized by a histogram. Data displays the summarized amount of sequence found for each similarity (number of path in the node). **(C)** Fraction of core, soft and private nodes in the whole graph (right) and sequence in each genome (middle) and total sequence (left).

comparable proportions of private and soft-core pangenome content. However, accessions such as 9905 and 22003 stand out with unusually high levels of private sequence. At the same time, these accessions exhibit notably low soft-core content, suggesting that a significant portion of their unique sequence may, in fact, be shared with other accessions but remains unaligned due to structural complexity, assembly differences, or divergence beyond alignment thresholds. This observation highlights the limitations of alignment-based classification and suggests that private sequence content can be inflated in genomes with lower-quality alignments or increased structural variation.

One possible explanation is that these private sequences originate from centromeric or other highly repetitive regions, which are inherently difficult to assemble and align. Given the sequencing technologies used, it is likely that these regions were only partially captured

or incompletely assembled, limiting accurate placement within the graph. This technical limitation may lead to the erroneous classification of repetitive or poorly aligned sequences as private, when in fact they may be shared across multiple accessions but remain unresolved in the current dataset.

The distribution of shared sequence across the 28 accessions reveals a U-shaped pattern (Figure 3.4B, with most sequences being either private (found in only one sample) or core (present in all samples). Intermediate levels of similarity are less frequent, indicating that much of the genome is either highly conserved or highly individual-specific. This distribution suggests a stable genomic background common to all samples, alongside a considerable amount of accession-specific content. The large proportion of private sequences may reflect true biological variation, unalignable regions, such as centromeres, or potential assembly artifacts, particularly in repetitive or structurally complex regions.

Regional similarity

We compared complexity patterns across accessions and observed that nearly all exhibit a consistent structure of alternating low- and high-complexity regions. Chromosomes are primarily composed of sequence shared across all accessions, particularly along the chromosome arms, which are known to be gene-rich. In contrast, sequence similarity decreases sharply near centromeric regions. Although the exact positions of centromeres vary between chromosomes, they are consistently identifiable across accessions. Minor positional shifts are due assembly length introduced by the use of CLR sequencing technology.

Notably, even pericentromeric regions, located approximately 2–4 Mbp from the centromere, show an early decline in similarity. This pattern suggests a more dynamic and structurally variable genomic landscape, likely influenced by increased transposable element (TE) activity and rearrangements that disrupt conserved gene content or synteny. These complexity patterns are visually marked with black arrows in the corresponding figures for clarity (Figure 3.5).

Arabidopsis thaliana saturation

In our dataset, we observed that the pangenome did not appear to be saturating even after the inclusion of 28 genomes, which indicates that new sequences are continuously being added to the pangenome (Figure 3.6A). Interestingly, the behavior of the core genome, as opposed to the soft and private sequences, appears to plateau after several accessions and remains relatively stable thereafter. These findings suggest that the impact of new sequences on the pangenome is largely driven by the addition of private sequence.

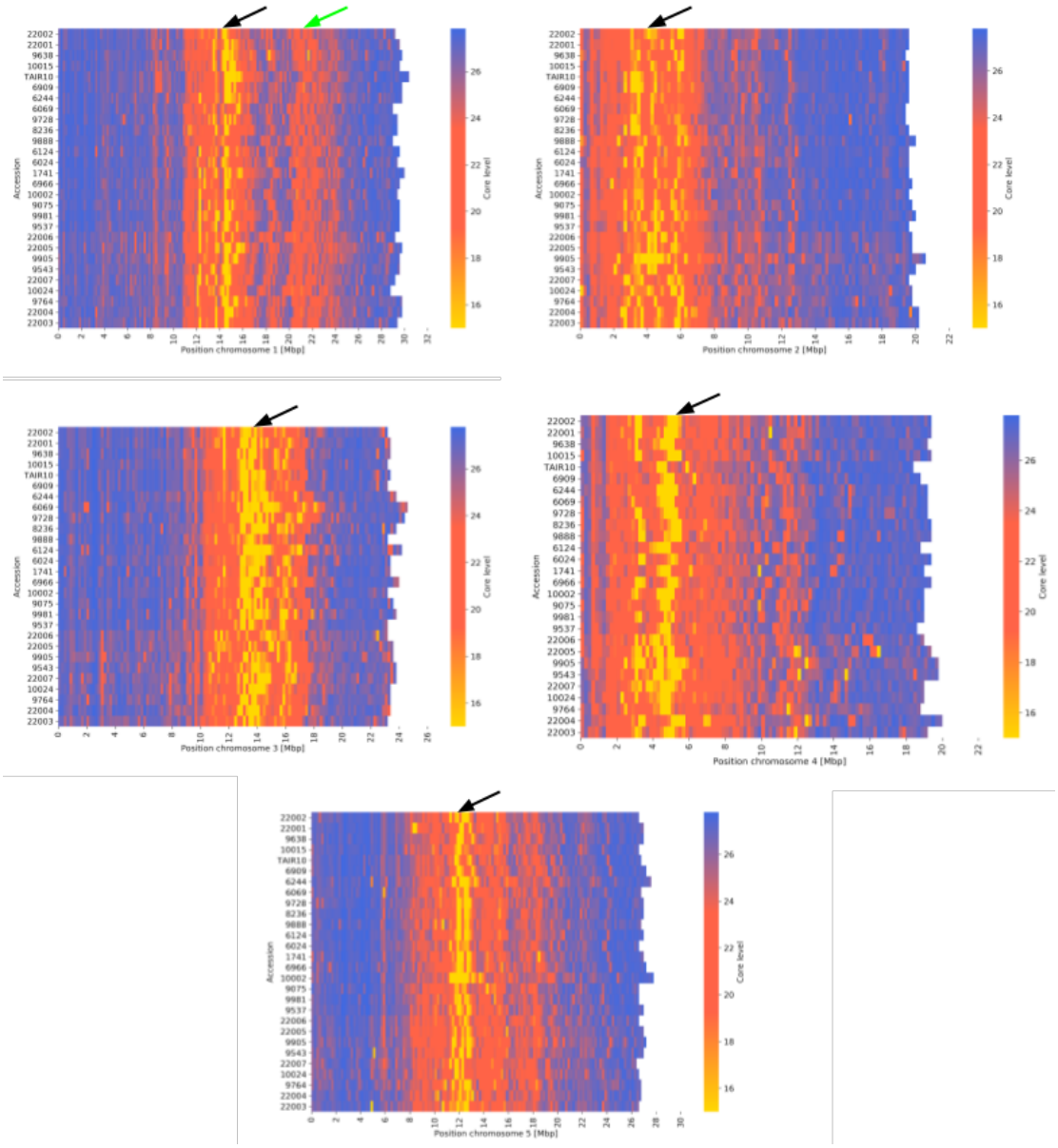
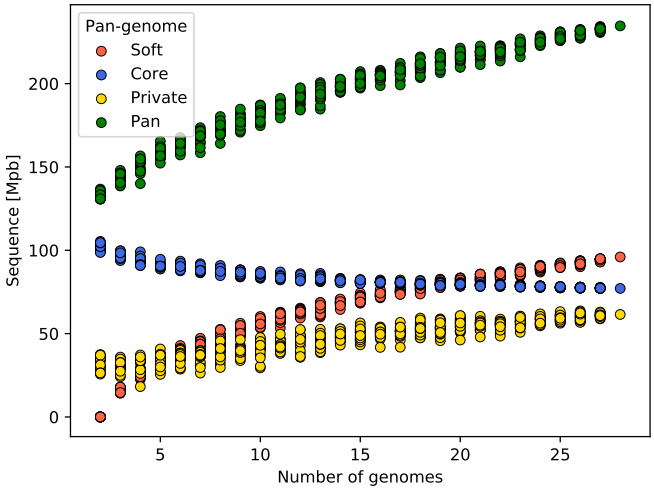
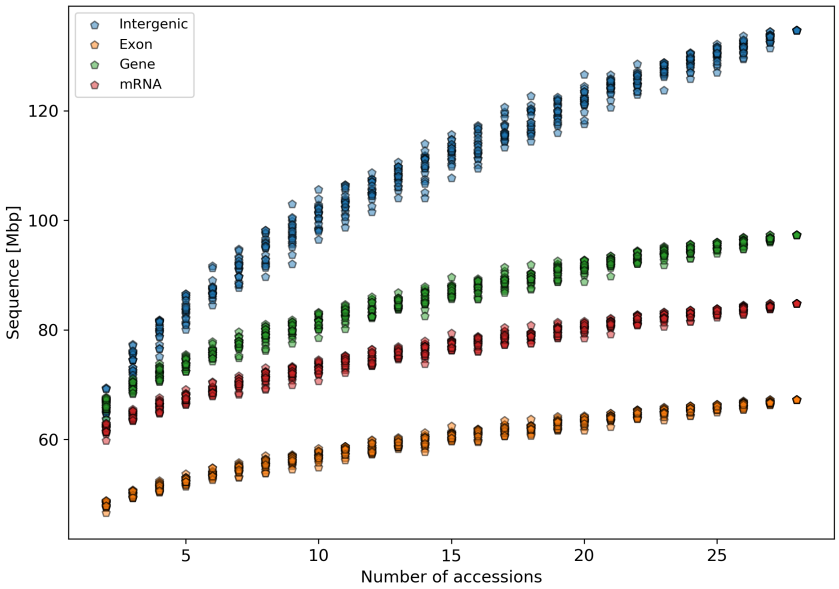


Fig. 3.5 **Sliding window similarity plots.**
These heatmaps display the similarity level along the chromosome for each of the 5 chromosomes, with the x-axis showing the chromosome position and the y-axis showing the accession used in this analysis. Windows are 500 kbp in size and are represented as color-coded single value (arithmetic mean). Blue indicates high similarity and yellow indicates low similarity. Black arrows mark centromeres and green arrows indicate chromosome fusion sites. For better visualization, similarity values are cut by 15 and every window with lower values is shown as yellow.

Expanding on the previous analysis, we also included annotation data. By using the same approach, we can determine how much new information is added to each category of genes



(A) Saturation analysis based on 20x bootstrapping based on different combinations of path in the graph.



(B) Saturation on intersection between the genome graph and the annotation data set. Exons, genes, mRNA and intergenic regions are shown in this plot.

Fig. 3.6 ***Arabidopsis thaliana* saturation.**

and their respective parts, such as introns and exons. When overlapping the analysis with annotation data, we have found that most of the new sequence accumulates in the intergenic regions, which are located between genes. In contrast, the genes, mRNA, and exon parts of the genome seem to accumulate new sequences much more slowly than intergenic regions, and show much less variation within each bootstrapping approach (Figure 3.6B).

Notably, the increase ratio of the intron, mRNA, and exon sequences seems to be the same, indicating that these categories have the same limitations/mechanism in terms of accumulating new sequences.

3.3.3 Bubble statistics

To compare our graph in terms of variation with other approaches, we decomposed it into distinct bubble structures. Bubbles are defined by distinct start and end positions, which we use as unique identifiers to accumulate statistics across our graph set. Each bubble contains at least two traversals from the start to the end node and is not connected to other bubbles, except at these boundary nodes. Additionally, bubbles can be nested within larger bubbles if they are fully enclosed. Most of the subsequent analyses were conducted using SV bubbles, containing at least one allele/traversal longer than 14 base pairs. Across our five chromosome graphs, we detected a total of 5,468,158 bubbles. The median bubble size was 1 bp, while the average bubble size was 32.1 bp. The largest bubble observed reached a size of 8,253,189 bp.

When focusing on SV-related bubbles, we observed a general (exponential) decrease in bubble size, with most bubbles being relatively small (Figure 3.7C). An interesting deviation from this trend is a local peak at approximately 5,000 base pairs, which likely reflects a specific structural feature or recurring larger-scale variant.

Overall, most SV bubbles are traversed by all samples in the graph (Figure 3.7B). This is expected, given the high-resolution chromosome-scale assemblies and the relatively uniform distribution of genome complexity in the dataset. Nevertheless, bubbles traversed by fewer samples are also common, forming a U-shaped distribution. Particularly notable is the presence of bubbles traversed by only a single path, indicating the existence of sample-specific variation.

As shown in Figure 3.7A, most bubbles consist of exactly two distinct traversals. Furthermore, the number of bubbles decreases rapidly with an increasing number of traversals. It is important to note that traversals can originate from different samples or from repeated segments within the same sample. While bubbles with more than 25 distinct traversals are rare, they may occur in a complex and collapsed graph context.

Finally, we analyzed the size differences between the alternative traversals (alleles) within each bubble. The resulting distribution shows a U-shape, reflecting two dominant classes of bubble structure (Figure 3.7D). Bubbles with a size ratio of 1.0, exhibiting a peak on the right, are particularly prominent, representing cases where the smallest and largest alleles are approximately the same size. On the left, we observe InDels of varying types, including insertions where one allele is zero base pairs in length. These may also include larger insertions or cases where one allele represents an extended insertion and the other a nearly

absent deletion. Such cases may explain the plateau observed near the 0.0 ratio. In graphs, indels often lack clearly defined boundaries, sometimes introducing sequence at the deletion site, which may not perfectly reflect the biological truth. Nevertheless, these structures are identifiable with our analysis.

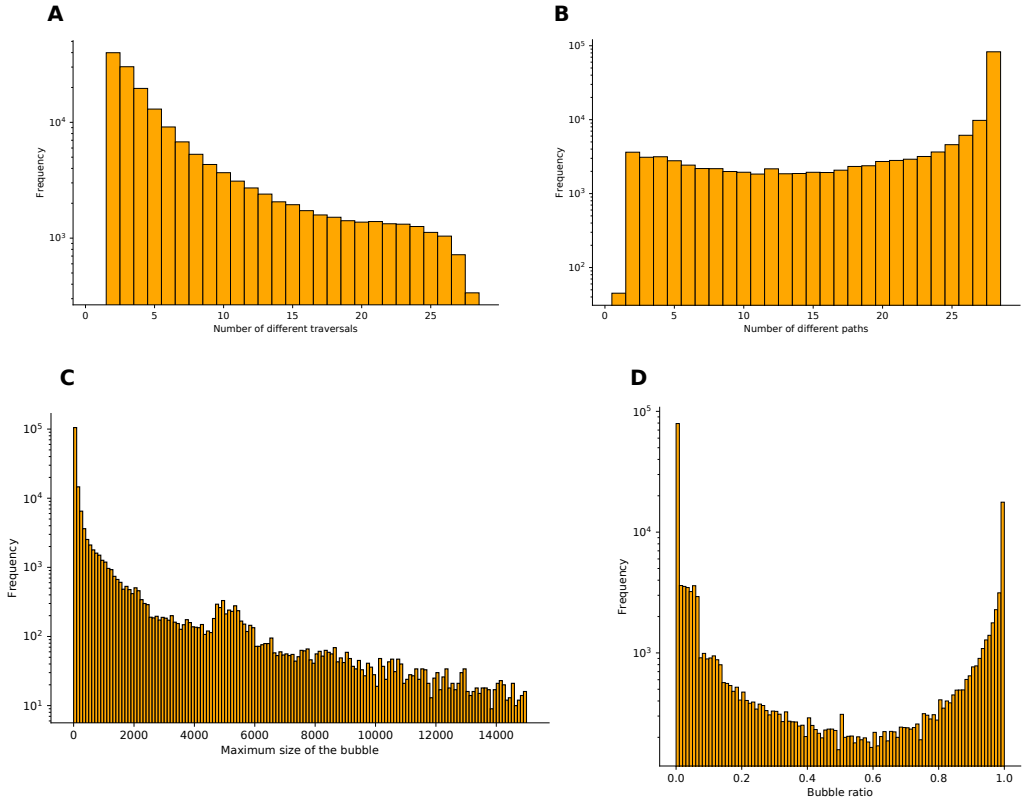


Fig. 3.7 **Statistics of variation from genome graphs.**
(A) Number of different traversals (alleles) within each bubble. **(B)** Number of different accessions in each bubble. **(C)** Histogram displaying the maximum bubble size for each bubble found in the graphs. **(D)** Bubble ratio (minimum allele size divided by maximum allele) histogram.

3.3.4 Structural variation comparison with Pannagram

Obvious large-scale rearrangements aside, a comprehensive characterization of structural variants (SVs, by which we mean any alteration that causes variation in length, orientation, or local context of sequence) remains difficult. While SVs, along with SNPs, can be identified in genome alignments, the characterization of SVs is a fundamentally different problem from SNP-calling. The latter can be viewed as a technical issue—how to distinguish single-nucleotide polymorphisms from sequencing errors—but SV-calling is challenging even with

flawless chromosomal sequences. The reason is that the SVs identified between genomes depend on the alignment method and parameters used, and there is no obvious ground truth. Given these uncertainties, we pursued two complementary approaches. The resulting object is computationally efficient for genotyping, but is also highly complex, with neither nodes nor bubbles having an obvious biological interpretation.

Comparing the SVs identified by the two conceptually different approaches was not straightforward. SVs identified by Pannagram were typically covered by PGGB variants, which included nearly twice as much sequences, especially in highly polymorphic pericentromeric regions (Figure 3.8). The overall overlap between the two methods was 52% without any prior filtering. In this unfiltered comparison, Pannagram contributed almost no additional sequences classified as SVs. In contrast, the graph-based approach annotated approximately 40% more sequences as SVs (Figure 3.8A). Further analysis revealed that many of the SVs identified by the graph spanned regions with poor or no reliable alignment, often extending several kilobases. This was especially common in pericentromeric and centromeric regions (Figure 3.8C). By removing SVs larger than 100 kbp from the comparison, the overall overlap increased, and both graph-specific and Pannagram-specific SVs added modest amounts of sequences (Figure 3.8A).

Chromosome-level analysis (Figure 3.8C) showed that the ratio between shared, graph-specific, and Pannagram-specific SVs was nearly consistent across all chromosomes. This indicates that the differences between the two methods are systematic rather than chromosome-specific.

When analysing the genomic distribution of SVs, we observed that SV counts were generally similar on chromosome arms. However, the graph still identified approximately 40% more sequence in these regions. The largest discrepancies were found in the pericentromeric regions, where the majority of SVs were located. Both graph and Pannagram detected increased numbers of SVs in these regions, although the rise in Pannagram was moderate. In the centromeric regions, Pannagram internally filtered out most SVs, resulting in SV calls only from the graph-based approach. The size and structure of the pericentromeric regions varied between chromosomes, which did not significantly affect the overall statistics but is visually present in our analysis (Figure 3.8C). Additionally, Pannagram reported a few unique SVs in these regions, though their total contribution was low.

A trivial reason for this difference is that Pannagram masked centromeric regions full of tandem repeat arrays , but we also identified several less obvious causes. Details and examples are shown in Figures 3.9–3.12.

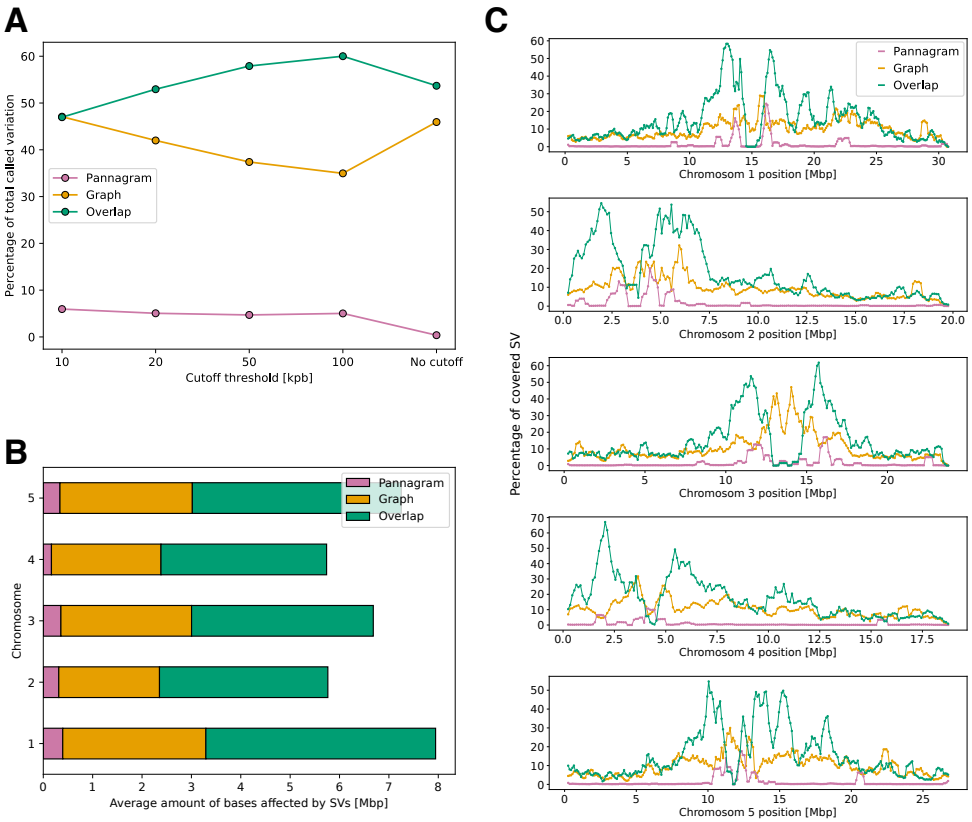


Fig. 3.8 **Comparing SVs from Pannagram and PGGB graphs.** **(A)** Scatter plot of overlap and method-specific SVs as a function of eliminating SVs above a certain length-cutoff. Overall, the overlap between the two methods is 50%, but the overlap can be increased by removing large SVs (demonstrating that disagreement is disproportionally due to large SVs. **(B)** Comparison of Pannagram and graph-based SVs across chromosomes (average per accession), demonstrating that there are no major differences between chromosomes. SVs shorter than 15 bp were not included in this figure. **(C)** Position of overlapping and method-specific SVs for each chromosome of accession 6909 (Col-0). Large discrepancies are more pronounced close to the centromeres (there is no overlap inside centromeres, as these are masked by Pannagram). Each dot represents a 100 kb window, using a moving average of five windows.

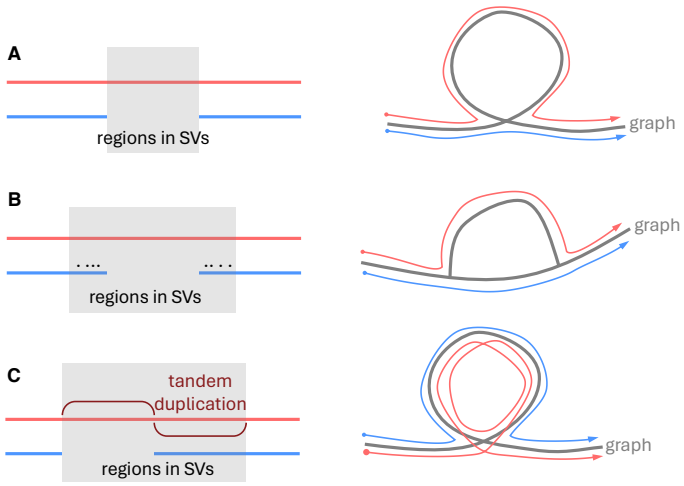


Fig. 3.9 **Cartoons illustrating cases where graph SVs are longer than Pannagram SVs.** **(A)** Two genomes (red and blue) differ by a single simple SV (Figure 3.10), which can be represented as a gap in the alignment, or a loop in the graph. Pannagram and the graph give the same result and the length of SV (indicated in grey) is identical. **(B)** However, if SNPs, represented by dots, are linked to the SV, causing imperfect alignment in the flanking regions around the SV, PGGB may recognize longer haplotypes, resulting in an arrangement that resembles a hat. In this case, the graph SV is not merely a presence-absence variant, but a complex SV with two alleles. The entire region affected (in grey) is longer than the SV recognized by Pannagram (still the same as in A). **(C)** When the SV is formed by a tandem duplication, the graph representation of the SV is topologically similar to scenario A, but the SV covers both the original sequence and its duplicated copy (grey region), while the SV identified by Pannagram is still the same as in A.

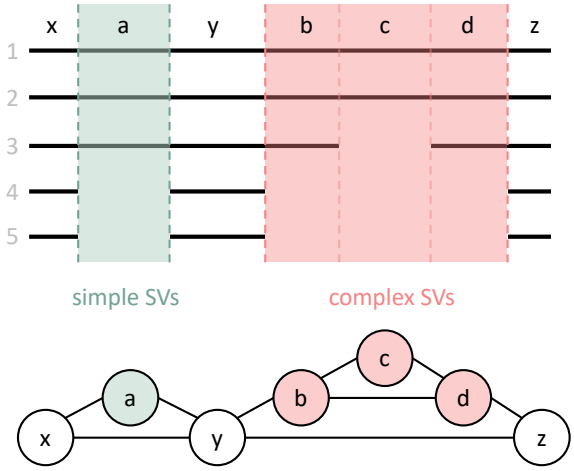


Fig. 3.10 **Simple and complex length variants.** Cartoons illustrating our classification of length variants into simple and complex structural variants (SVs) in the whole-genome alignment and graph representations.

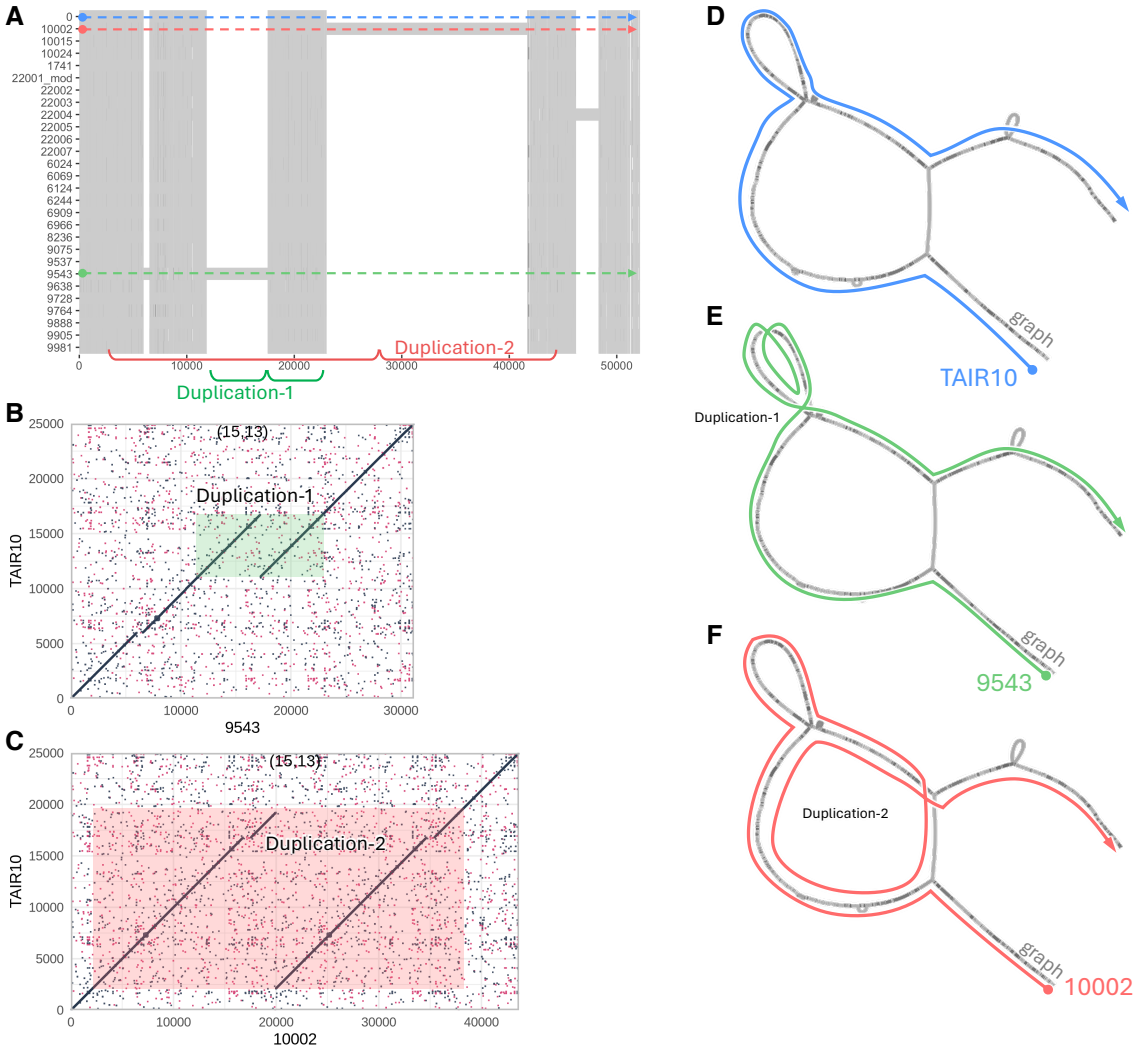


Fig. 3.11 **An example of how Pannagram and the PGGB graph each handle closely linked duplications.** The region depicted corresponds to coordinates 285,000-310,000 bp on chromosome 1 in accession 1741. The Pannagram alignment (panel **A**) identifies four simple SVs, with the two longest ones being due to duplicated sequences in accession 9543 (panel **B**) and accession 10002 (panel **C**). The PGGB graph representation of this region is shown on the right, along with paths corresponding to three different haplotypes. Panel **D** shows the path taken by accession 0 (TAIR10), which carries the majority haplotype. Panel **E** shows the path of accession 9543, which carries a duplication, and hence goes around the small loop twice. Panel **F** shows the path of accession 10002, which has the longest duplication, and hence goes around the big central loop twice. Thus, while Pannagram identifies four simple SVs (the longest one being 18.6 kb long), the PGGB graph SVs involve all accessions and cover almost the entire region shown. Note that, as in the cartoon example (Figure 3.9), similar PGGB graph topologies may result from very different types of sequence differences between accessions.

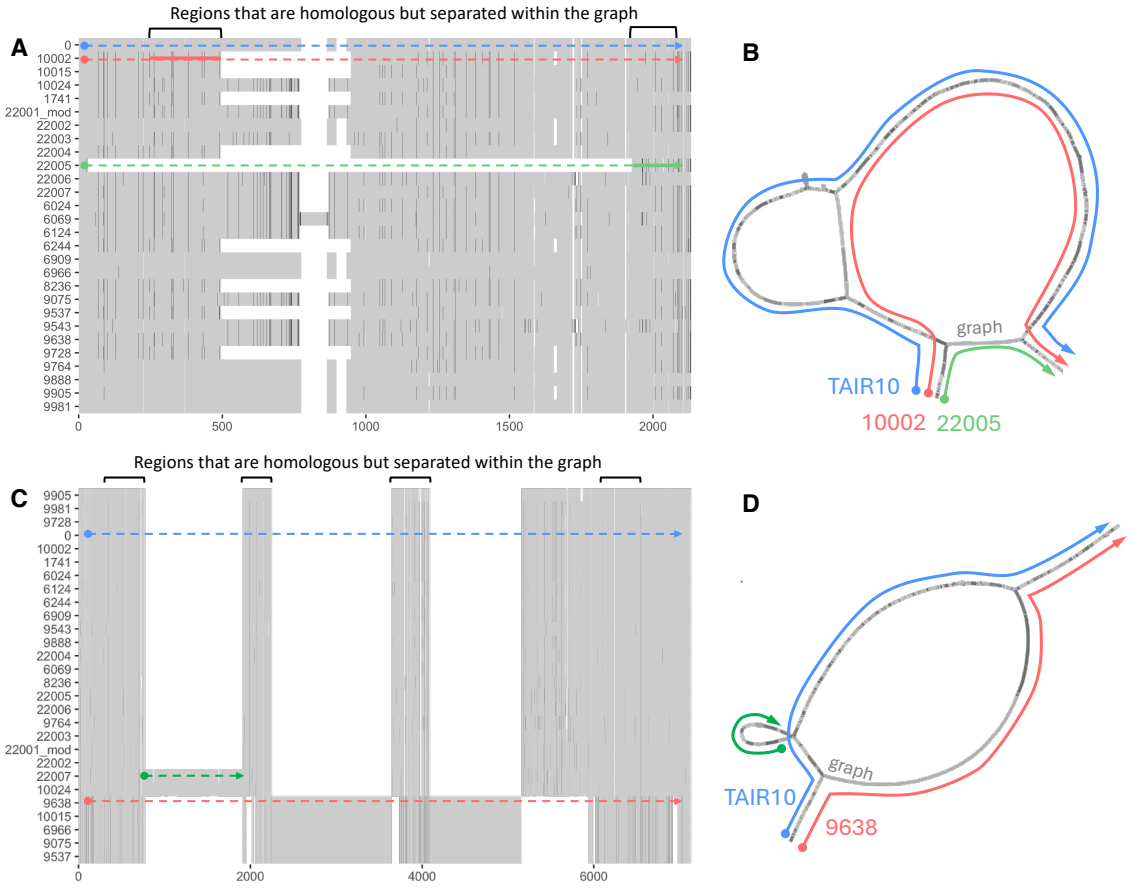


Fig. 3.12 **Another example of how Pannagram and the PGGB graph each handle regions that are difficult to align.** **(A)** Pannagram alignment of a highly polymorphic region corresponding to coordinates 21,956,400-21,958,060 bp on chromosome 1 in accession 1741. Pannagram identifies a complex SV covering most of the region. **(B)** The PGGB graph also recognizes these SVs, but merges them with flanking SNP variation, resulting in two nested hat-like structures (*cf.* Figure 3.9B). As a result, the sequence covered by SVs is longer. **(C)** Pannagram alignment of the region corresponding to coordinates 1,183,130-1,186,590 bp on chromosome 1 in accession 0 (TAIR10). Pannagram identifies several, mostly simple SVs separated by short alignable regions. **(D)** The PGGB graph does not align these regions, and merges most variants into two longer haplotypes. In this case as well, the graph SVs cover more sequence than the Pannagram SVs.

3.4 Discussion

The clear separation of accessions into two major clades based on Mash distance highlights the strong population structure within the *Arabidopsis* dataset, reflecting both geographic origin and historical lineage. The clustering of European and Asian samples, especially the tight grouping of Chinese accessions, is consistent with expectations from earlier genomic studies [3]. The identification of highly similar pairs such as 22005/22006 and 6069/6244 confirms the presence of near-identical genotypes within the dataset, likely due to sampling from closely related populations.

The second clade, consisting largely of African, Madeiran, and relict accessions, reinforces the distinct evolutionary trajectories of these groups. Their separation from the core European/Asian cluster supports previous hypotheses about their status as deeply diverged or ancestral lineages within *Arabidopsis*. Interestingly, the presence of clear subclades within this group, particularly among relict and Madeiran samples, suggests a finer-scale structure that merits further phylogeographic analysis.

While Mash is well-suited for identifying global sequence similarity, its limitations in resolving complex genomic rearrangements prompted the use of SyRI and DeepVariant. That both of these complementary methods reproduced similar clustering patterns adds confidence to the inferred genetic structure and highlights the robustness of our observations across distinct methods. Together, these results suggest that both point mutations and large-scale structural differences contribute to population differentiation in *Arabidopsis*, and that integrative approaches combining alignment-free and alignment-based tools are essential for a comprehensive view of genome variation.

The PCA results reinforce the population structure identified through distance analyses, confirming clear clustering of Central European accessions and greater dispersion among non-European genomes. This pattern likely reflects reduced genetic diversity in European accessions due to shared ancestry or demographic history, while broader variance in accessions like 22001, 22002, 10015, and 6938 may indicate ancestral divergence or geographic isolation. The close proximity of TAIR10 and 6909 further validates the method, highlighting PCA's effectiveness in capturing both broad and fine-scale genomic relationships. It is important to note that our PCA approach included all nodes without any pre-filtering. Removing high-depth nodes could potentially improve population structure resolution, as these often represent repetitive or less-informative regions. However, our method currently assigns equal weight to all nodes, regardless of size or complexity. As a result, large insertions and small SNPs contribute equally to the analysis, despite representing different scales of variation.

To better capture structural and contextual variation, we also suggest exploring the use of edges. Unlike nodes, edges not only carry sequence information (via their connected nodes)

but also encode local topology and positional context within the graph. This could offer a richer representation of genomic variation and might be a promising direction for future analyses.

Furthermore, the binary presence/absence matrix derived from graph features could potentially serve as a kinship matrix for downstream applications such as GWAS, provided a sufficient number of accessions are included in the graph. However, it is important to acknowledge that current SNP-based GWAS approaches, which rely on curated and polished variant datasets, offer higher resolution and reliability. Therefore, future efforts should focus on developing methods to polish and filter graph-derived data to extract high-confidence variation that is suitable for quantitative trait analysis.

Our analysis demonstrates the utility of genome graphs in capturing population-level variation and structural complexity in *Arabidopsis thaliana*. Over 60% of the sequence is shared among all accessions, highlighting a relatively conserved genome. However, this estimate excludes unresolved alignments, shared-but-non-core sequences, and many small variants. Private sequence content is often inflated in accessions lacking close relatives in the dataset, while accessions with nearby genetic neighbors show a higher proportion of 'soft' pangenome. With the current dataset, our selection remains relatively unique, as the original 1001 Genomes (1001G) project includes a large number of genetically closely related accessions. Nevertheless, careful sample selection is crucial for understanding the full extent of differences between accessions within the pangenome space. The close similarity between TAIR10 and accession 6909 confirms the completeness of our assemblies, while differences in centromeric and repeat-rich regions likely reflect sequencing limitations. CLR-based assemblies, though widely used, often fail to resolve highly repetitive sequences, leading to unaligned or misclassified regions. Modern long-read technologies like HiFi and Nanopore offer greater resolution and throughput, which will improve representation of such complex loci in future pangenome graphs.

At the chromosomal level, regions of low similarity consistently align with centromeres and other repetitive elements, reflecting both biological divergence and technical constraints in graph construction. Some low-complexity regions correspond to known fusion events derived from *Arabidopsis lyrata*, though others are harder to detect without prior knowledge [47, 86, 56]. Combining similarity data with depth information helps distinguish biologically meaningful variation from assembly artifacts, regions with both low similarity and high depth are likely hotspots of structural complexity and warrant targeted investigation.

Pangenome saturation analysis further reveals that while the core genome quickly stabilizes, soft and private regions continue to expand linearly with the addition of new accessions. Much of this new sequence arises from intergenic regions, which can accumulate varia-

tion without functional impact. In contrast, coding regions remain highly constrained, as mutations in essential genes are typically purged from the population. This suggests that functional constraints shape the structure and evolutionary dynamics of the *Arabidopsis* genome, with new variation largely restricted to non-coding regions.

This finding highlights the importance of studying not only the sequence composition of an organism but also the annotation data, as both can provide valuable insights into the underlying genetic mechanisms of an organism. Overall, the saturation analysis approach provides a powerful tool for exploring the complexity of an organism's genome and for identifying the main drivers behind the growth of its pangenome.

While similarity scores across accessions offer a broad overview of genome conservation, bubble structures in the graph provide a more detailed and localized perspective on genomic complexity. Most bubbles in our analysis were small and low in complexity, typically representing SNPs, simple insertions or deletions.

In more complex regions, bubbles become highly nested and structurally difficult to interpret. These regions often contain overlapping paths, long insertions, and alternate paths with shared nodes, complicating interpretation. While some of these structures reflect true biological variation, others may result from technical artifacts during graph construction or sequence alignment errors. Distinguishing between these cases remains a challenge, especially in the absence of supporting functional annotations or experimental validation.

The observed exponential decline in bubble size, along with the local peak at approximately 5,000 bp, suggests the presence of recurring structural patterns or conserved long insertions, possibly associated with specific functional or repetitive elements. Another possible explanation is the presence of technical artifacts, particularly during the normalization step of `smoothxg`, where graph regions are re-aligned to produce a base-resolution graph structure. These re-alignment windows are typically around 5 kb in size, depending on node length and count. If a region fails to align within one window, it may be incorrectly left alone and only successfully aligned in the following window, potentially creating artificial fragmentation and inflated private SVs.

The traversal profiles of SV bubbles further highlight key features of genome diversity. The predominance of bubbles traversed by all accessions indicates shared variation, consistent with the high-quality, chromosome-scale assemblies used. However, the U-shaped distribution, particularly the presence of bubbles unique to single samples, reveals the extent of accession-specific events, which may arise from unique duplications, TE insertions, or assembly artifacts.

Structurally, most bubbles consist of exactly two traversals, underscoring the dominance of biallelic variation within the population. Bubbles with many traversals are rare and often

more difficult to interpret biologically, as they may arise from complex or repetitive regions. Given that genome graphs are cyclic by nature and tend to collapse distant but similar regions in areas of high repetitiveness, it is possible that densely connected subgraphs are formed without clear biological relevance. Depending on parameter settings, such collapsed regions may simply represent common sequence motifs, e.g., domains shared across many proteins, without conveying functional or evolutionary significance beyond local sequence similarity.

Ratios are distributed as expected, but the plateau near a size ratio of 0.0 may reflect structural complexities in how insertions and deletions are represented in the graph, especially in regions where alignment boundaries are ambiguous (Figure 3.12). This pattern suggests that some insertions may span poorly resolved regions or include additional sequence context, making it difficult to cleanly define the variant boundaries and potentially inflating the apparent allele size imbalance.

Overall, bubble statistics offer a powerful means of capturing and quantifying structural variation in pangenome graphs. While small biallelic events dominate, the identification of sample-specific and complex multi-allelic bubbles provides a window into accession-level uniqueness and graph complexity. Further refinement of classification methods, particularly with respect to traversal diversity and size asymmetry, will be key to improving interpretation and comparative analyses across tools and datasets.

In the context of comparing graph-based SV analysis to results of Pannagram, the presence of physically distant but closely related sequences (*e.g.*, reflecting recent TE activity) can lead to large loops in the PGGB graph that do not reflect actual SVs. Masking repetitive sequences will reduce this problem, but requires good repeat annotation—and would also make it impossible to study genome-variation comprehensively. Even in the case of tandem duplications, the graph combines duplicated sequences into a single node and hence counts all these sequences as part of SVs, even if not all of them are variable. PGGB and Pannagram rely on different alignment parameters. The *A. thaliana* genome contains many highly diverged regions [21], and these tend to be treated by PGGB as long SVs, whereas Pannagram often finds short alignments, resulting in local clusters of shorter variants. Whether Pannagram or PGGB results are more biologically relevant ultimately depends on the question and the cause of high divergence. The differences between two algorithms designed to do different things should not be interpreted as bias, and that, as noted above, there is no ground truth.

Chapter 4

Graph GWAS – Gfa2bin and applications

This chapter is based on the publication *Gfa2bin enables graph-based GWAS by converting genome graphs to pangenomic genotypes*, available on bioRXiv at https://doi.org/10.1101/2024.12.05.626966 and currently in the process of publishing in a journal.

4.1 Introduction

The advent of long-read sequencing technologies has revolutionized the field of genomics, enabling the generation of more comprehensive and accurate genome information from multiple members of a species at a reasonable cost. These long-read datasets can be assembled into complete genomes, enabling more accurate determination of genome differences than was possible with short-read data [69, 85, 17]. The transition to long-read sequencing and complete genomes is overcoming the limitations of fragmented and incomplete representations, and is beginning to provide a more comprehensive view of genomic variation, including better haplotype information. Variation graphs have emerged as a promising alternative to traditional linear reference genomes, which often fail to capture the full spectrum of genetic variation present within and across populations or species [81].

By encoding non-reference alleles that include structural variants and complex genomic features, variation graphs offer a less biased representation of genetic diversity compared to a collection of linear genomes that have been aligned to a single reference genome [41, 39, 53]. This graph-based representation facilitates a wide range of analyses, such as genotyping, variant calling, and population genetics studies across diverse organisms [130, 52, 84].

The emergence of variation graphs presents an exciting opportunity to more fully exploit genomic data. As stated above, graphs offer a more comprehensive depiction of genetic diversity by themselves, but can also be used as a variation-aware reference for aligning

short- or long-read data sets. New approaches have been implemented for sequence-to-graph alignments, increasing mapping accuracy and sensitivity in complex regions [41]. Such an advanced data structure, either directly or with additional alignment, enables principled genome-wide association studies (GWAS) that go considerably beyond single nucleotide polymorphisms (SNPs) or variants linked to SNPs, thereby furthering the understanding of the entire spectrum of genetic differences responsible for phenotypic variation in a population.

It is intuitive that integrating variation graphs with GWAS should allow for a more thorough exploration of the genotype-phenotype map, and thus support the discovery of associations between DNA sequence variants and traits of interest that are not detectable with SNP-centric analyses. Detailed genotype information coupled with the capacity to represent a wide range of genomic features in variation graphs therefore promises to offer insights that were previously not easily accessible, such as understanding the functional consequences of multi-allelic and complex structural variants.

Conventional GWAS approaches typically rely on genetic variants inferred with the help of linear reference genomes to identify trait associations. This approach has several limitations, particularly when there is a high level of sequence differences between the reference genome and the samples being analyzed. In particular, this can lead to reference bias and the failure to identify causal genetic variations. A powerful alternative is being offered by genome graphs. So far, these were often constructed by anchoring them to a linear reference genome and incorporating known variants identified from reference-based methods, following a paradigm established already 15 years ago [122]. For example, Liu and colleagues derived a genome graph for soybeans (*Glycine max*) by detecting structural variants with the MUMmer toolkit using a single reference, and then genotyping almost 3,000 soybean accessions with the vg toolkit [84]. The efficacy of this approach was demonstrated through the identification of a specific structural variant on chromosome 15 that had not been previously linked to seed luster variation by GWAS or other means. Similarly, He and colleagues developed a pangenome graph for *Setaria italica* by aligning 112 genomes to a reference and identifying variants with Syri [43, 50]. Nearly 2,000 *S. italica* accessions were then genotyped with this graph and the variants were used in GWAS on 68 traits. Zhou et al used a most complex two-step approach to create a tomato genome graph [150]. Initially, they constructed a structural variation graph using HiFi long reads aligned against a reference genome. This graph was expanded by integrating additional variants identified from 706 tomato accessions using DeepVariant for calling indels and SNPs [113], and Paragraph for SV genotyping [16]. Unlike the other two studies, this method integrated all variants into a unified genome graph, capturing a broader spectrum of genetic diversity and missing heritability [150]. Impressive

as they were, these studies still had inherent limitations such as relying on reference-based approaches and the potential introduction of bias through variant representation in the VCF format. Additionally, the accuracy of genotyping using short-read alignment often falls short of the precision required for comprehensive variant detection [26].

As alternatives to graph-based approaches, several methods have emerged that completely eliminate the need for linear references. DBGWAS [61, 60] is one such method, using an alignment-free, k-mer based approach with compact de Bruijn graphs (cDBG). These represent overlaps between k-mers, removing redundancy, and capturing genetic variations in nodes and edges. Another approach, k-mer GWAS [49, 137, 42], determines the presence or absence of k-mers without relying on reference genomes. To avoid excessive computation for association of all k-mers, which are much more numerous that SNPs, Voichek and Weigel implemented a prefiltering step that greatly reduces the number of required association tests [137]. While elegant, a major disadvantage of k-mer-based methods is that they are very sensitive to sequencing errors, requiring careful filtering. Additionally, it is difficult to interpret the results without broader genomic context for the detected associations. To facilitate the incorporation of variation graphs into existing GWAS workflows, we created the tool `gfa2bin`, which converts genome graphs from GFA (Graphical Fragment Assembly) format into a graph-based genotype matrix suitable for GWAS analyses. For seamless integration with existing bioinformatics pipelines, we offer outputs in two different formats: the widely used PLINK format and the more versatile BIMBAM format [114, 124], and we provide a complete Snakemake [100] pipeline of the GWAS alignment-based workflow in the Github repository of gfa2bin. Using *Arabidopsis thaliana* as an example, we show that GWAS with variation graphs compares favorably to traditional reference-based approaches, although the best results are obtained when combining multiple methods. It offers unique benefits such as mitigation of reference bias and providing genomic context on the graph structure.

4.2 Implementation

`Gfa2bin` converts data from genome graphs and/or sequence-to-graph alignments into predominantly binary formats, which can subsequently serve as inputs for traditional GWAS methods (Figure 4.1. It does not introduce any new genome-wide association methods itself, but rather provides an easy-to-use tool for converting and modifying genotype matrices related to graphs.

The tool is implemented in the programming language Rust, incorporating several Rust repositories to enhance performance and enable multithreading. An additional tool, `packing`, for reducing storage of coverage files was developed in parallel. `Packing` can be used for compression and/or normalisation of coverage information (Appendix B.2).

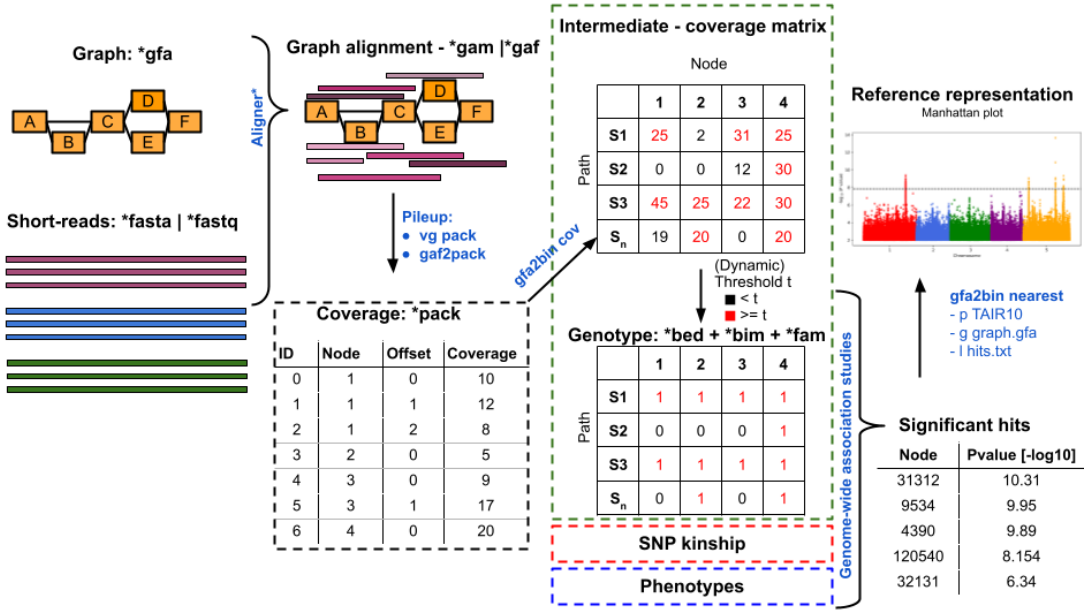


Fig. 4.1 **Schematic representation of the workflow in** `gfa2bin`.
Graphs can be converted to the binary matrix either by using the graph directly and reflecting features of the graph or from coverage information (shown here). Matrix can either be binary or value based and later be converted into multiple GWAs data types.

4.2.1 Data structures

The main data structure is a sample-by-feature matrix, holding bits or floating values dependent on output. We populate this matrix by either the data from the graph itself, or coverage information from a sequence-to-graph alignments.

The features are mostly nodes, but in specific cases, can also represent directed nodes or edges.

Edges, in comparison to nodes, indicate the genomic position and order of the sequences in the samples. Directed nodes can also be used as possible features, which represent nodes that enforce a specific direction, but mirror nodes-based representation in most cases (see GFA definition[1] for more information). Another feature, only available in our coverage method (`gfa2bin cov`), is the sequence (base pair) level.
The bit-wise matrix representation is realized using a low-level bit-vector implementation, reducing storage in comparison to a standard boolean vector representations. The matrix itself is feature-major (SNP-major in PLINK context), representing each feature as a distinct vector for fast and efficient computation and writing. SNP-major representations are generally better for computations on genotype level and output, but perform poorly for read-in operations, since almost all inputs come sample-major (e.g. one path per line or one sample-specific coverage vector).
It is important to note that the samples represented in our matrix, can be either (1) paths or samples of the graph itself (when running `gfa2bin graph`) or (2) the names of each aligned read set (`gfa2bin cov`). The second approach uses the graph as a variation-aware reference, which is reflected in the coverage itself and only indirectly in the resulting matrix.
Our bit-wise matrix is interchangeable with the PLINK (v1.9) representation of genotypes (BED and BIM format), which can be output by our conversion methods, or used as an input for post-processing or modifying (see below).

4.2.2 Data inputs

In `gfa2bin` our commands can be classified into two groups, (1) converting graph-related data to genotypes and (2) modifying or post-processing resulting PLINK files. In the following we listed all possible input formats.

Graph fragment assembly (GFA) format and Pan-SN spec

The GFA format is a tab-delimited text format for genome graphs, including their segments, edges (with optional overlap information) and sample information (e.g. paths and walks). In multiple iterations (version 1.1, 1.2 and 2.0), the format has been incrementally extended to cover other, specific features which can be represented by a graph structure.
The usage of Pan-SN spec assists to group multiple paths to a single sample, which can be helpful in the context if assemblies are of contig level because of organism complexity or assembly method. Other GFA construction tools use "walks" instead of paths in the context of pangenome graphs.

[1] https://gfa-spec.github.io/GFA-spec/GFA1.html

Graph coverage

The coverage obtained by sequence-to-graph alignments is normally calculated on a base level, summarizing the number of all mapped reads at certain position. Our tool has a broad range of possible inputs, including plain-text formats or highly compressed binary files.
Pack files are tab-separated files with all necessary information needed to track down the exact position of the covered sequence. The positional information is shown as node id ("node.id") and the offset ("off.set") that represented the relative position on the specific node. In addition, the overall position in the graph ("seq.pos") is stated. Nodes are numerically sorted by identifier, starting at the smallest at the top and ranging to the smallest at the bottom of the file. This file can easily be generated by several methods (see below) and offers the user a fast way to display the coverage and corresponding position.

Packing tool

The `packing` tool was designed during the work on the projects of this thesis to reduce the storage demands of `pack` files. One very straight-forward solution was to remove the redundant positional information at the beginning of each line, since it is the same for each coverage file derived from the same graph. Therefore, we created a simple index structure, only holding the "node" column of the original pack file (Appendix B.2). We can cross-compute the sequence position and offset using this data alone.
The coverage itself is also represented in a single vector, each value converted to 4 bytes (unsigned 32-bit integer), and then compressed with zstd. We also offer several methods to directly scale or threshold the input coverage information to other data formats. These methods are mainly used for pre-computation of coverage files, which can number in the ten-thousands in some experiments. The tool was developed in Rust and is available at https://github.com/MoinSebi/packing.

PLINK

PLINK, a widely used tool in bioinformatics, employs specific file formats for efficient genetic data analysis [114]. The primary formats are the PED and MAP files. The PED file encompasses genotype and phenotype data for each individual, detailing familial and individual identifiers alongside a comprehensive set of genotypes for all studied markers. The MAP file complements the PED file by providing detailed information about genetic markers, including their chromosomal positions and unique identifiers.
Additionally, to enhance processing speed, PLINK supports binary file formats: the BED file stores binary genotype data, the BIM file contains marker details paralleling the MAP file,

and the FAM file records demographic and phenotypic data similar to the initial sections of the PED file.
These file formats are crucial for conducting efficient genome-wide association studies and other genetic analyses, enabling researchers to handle large-scale data effectively and advance the field of genetic research.

4.2.3 Sequence-to-graph alignments

Short-read data

Although long-read sequencing made tremendous advances in the last few years, short-read data in GWAS remains valuable in direct comparison. Short-read technologies offer higher accuracy and throughput, making them reliable for variant calling. They are also cost-effective, allowing larger sample sizes within a budget. Furthermore, existing frameworks and analysis pipelines, are primarily based on short-read data. Additionally, the extensive use of short-read data in large-scale GWAS studies has resulted in established associations and reference panels. Overall, short-read data's accuracy, cost-effectiveness, compatibility, and established use make it still relevant and practical for GWAS.

Mapping to graphs

Mapping short-read data sets remains a challenging task and is an active area of development within the field of genome graphs. While several approaches have been published, internal testing have shown that these methods primarily rely on simple graph structures from tools like VG or Minigraph-Cactus (MC). Moreover, the examples provided in these studies were predominantly based on human genomes, which are less complex compared to other eukaryotic genomes such as plants.
Nevertheless, the obvious way to do it is by directly align the sequences to the graph, mostly to an index representation of all the sequences. Of particular interest are the mapping algorithms implemented in the VG framework, including `vg map`, `vg multimap` and `vg giraffe`. Those methods have shown to provide good results when working with linear graphs. Another well-established method is called GraphAligner, which is adapted to handle long-read data.

Mapping workaround with "super genome"

Following discussions with Erik Garrison, an alternative alignment approach was proposed, which involves using established aligners like BWA instead of graph-based methods. The suggested pipeline incorporates a workaround by utilizing a concatenated super-genome built with the same samples used to construct the graph. In the next step, a super genome is employed for read alignment, utilizing tools like BWA mem. Reads are aligned to the "best" location in the super-genome, mitigating reference bias by selecting the optimal alignment within the set of individuals. The resulting SAM files can be converted to GAM format using VG's `inject` subcommand, effectively connecting the graph and super-genome alignments (s. Figure 4.2).
However, it is important to note that such alignments will always be specific to each sample, preventing the mappings of multipath alignments between different samples. Additionally, this method is sensitive to the number of secondary alignments, which may result in an inflated coverage in specific regions unless specific measures are implemented to address this concern.

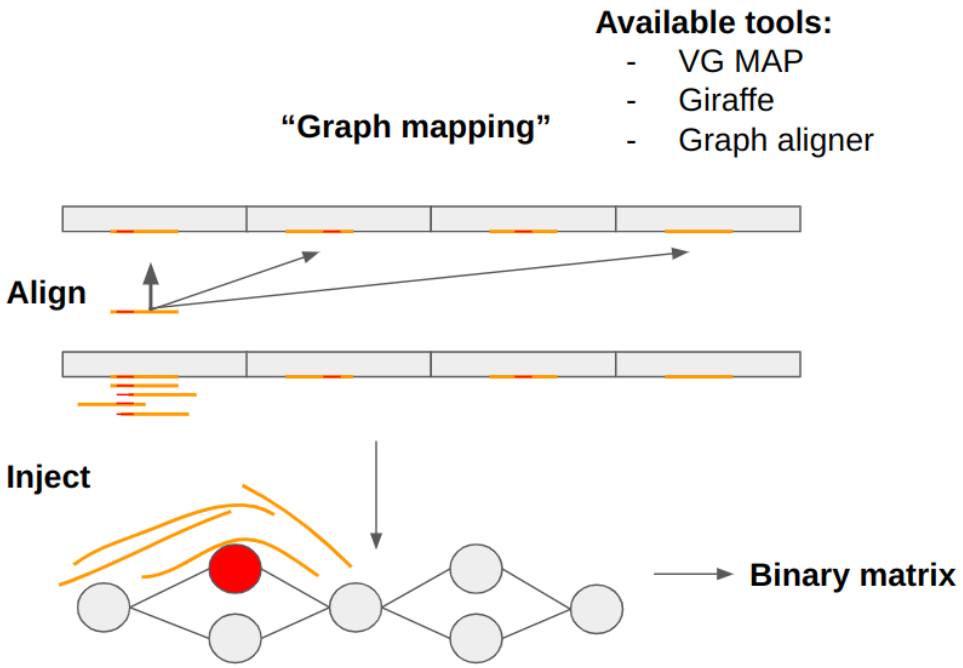


Fig. 4.2 **Schematic representation of the workflow for mapping sequence reads to genome graphs.** Reads are aligned with reference short-read aligner (e.g. BWA mem) to the concatenated super-genome. The resulting SAM/BAM file, can be converted to GAM with VG inject, since both approaches are based on the same genomes. Later, coverage is emitted by `vg pack`.

Snakemake workflow

We have added a Snakemake worflow that covers all steps used in our experiments (Figure 4.3). The workflow starts by validating input files and aligning the sequence reads to the graph. As stated above, we use a work-around since direct sequence-to-graph alignment can be unstable dependent on genome complexity and graph construction technique. The resulting linear alignments are then converted to the graph (inject), the coverage is summarized and everything is converted to a single PLINK file. This is later used in a GWAS, including the passed kinship matrix and phenotypes (validated in the first step). We report the (graph-)position and *P* values in a tab-separated table (output of GEMMA) and multiple plots from our custom plotting scripts.

4.2.4 Output data formats and copy number variation

This converter produces various output files, primarily in PLINK data formats, as they are widely used in established GWAS methods. The PLINK data format is based on a binary data structure. Therefore, the value-based matrix will be thresholded to obtain a presence-absence matrix, which can then be converted (s. Figure 4.1). This conversion and any other filtering can be performed using different subcommand flags. Additionally, it is possible to generate a BIM file that was originally used for imputed genotypes. This format assigns values ranging from 0 to 2, representing reference or alternative alleles (homozygotes) or any imputed value in-between. We can utilize this measure to accommodate copy number variations by scaling the maximum copy number to 2. Different scaling methods can be applied to reflect values between zero and two.

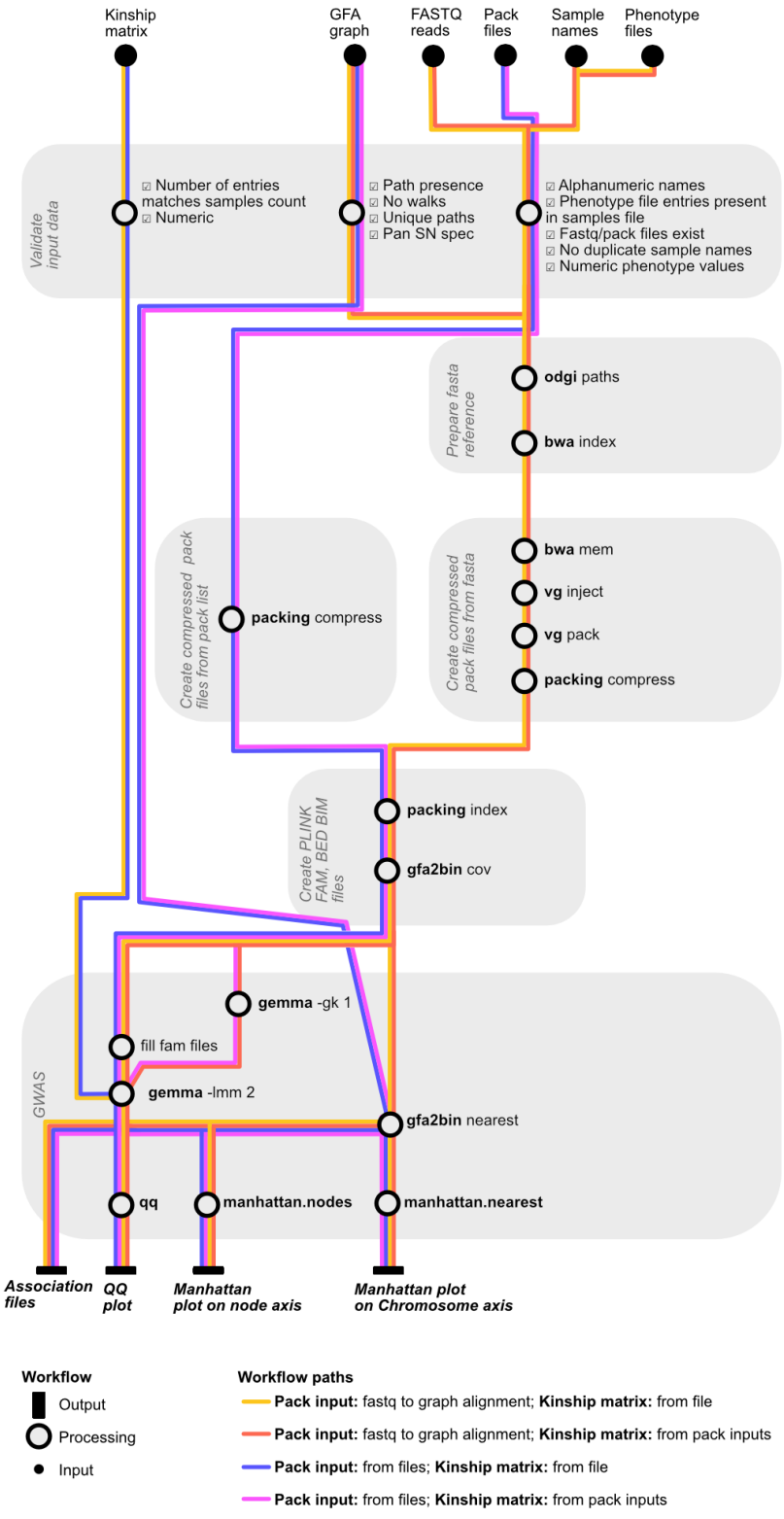


Fig. 4.3 **Diagram of the graph-based workflow as implemented in the provided pipeline.** Sequences can be aligned linearly to all genomes simultaneously, with linear alignments translated into graph terms using the graph structure itself. This conversion facilitates the calculation of coverage by the graph alignment. Alternatively, externally generated pack files can be processed.

4.2.5 Method overview

Here, we present a selection of commands of `gfa2bin`. This documentation is also available at the Github repository of `gfa2bin`[2], where it is kept up-to-date.

`Graph` - Converting from graphs directly

Convert a graph in gfa format to a PLINK or BIMBAM format. You are able to specify which feature (`-f`) you want to use to genotype. We support nodes (`1`), edges (`1+2+`), and directed nodes (`1+`). Paths can be merged to samples using the PanSN-spec, which is highly recommended. We count occurrences of each feature in each path/sample in the graph and use them as genotypes.

We are able to provide information about ploidy based on the PanSN-spec (PanSN must be used). In PLINK files, ploidy can easily be represented by `11`, `01`, `10`, `00`. In a BIM file, we use the average of both "scaled" haplotypes.

`Subpath` - Include neighboring information

In contrast to the "graph" subcommand, this subcommand is able to include neighboring information for a node and convert it. In general, we iterate over each node in the graph, extract the subpaths (+/- X steps away from the origin) traversing this node, and collect all subpaths. Subpath-groups (same subpath found in different paths) are then used as genotypes.

`Cov` - Using graph coverage

Convert coverage information from sequence-to-graph alignments to genotypes. Either can use plain-text pack files (also zstd compressed) directly or use one of the custom coverage file formats from packing repository as input. The packing repository helps to reduce storage and can perform pre-processing on sample level. Comparable to the graph subcommand, we offer additional normalization on sample-level.

`Find` - Extract the exact position of a genotype

Extract the genomic positions of a given list of genotypes. This methods outputs a BED file of those paths, where such genotypes can be found. If users might need more than just the exact position, additional `-length` information can be added, which will return also the region "neighboring" the node, adding the additional length to each site.

[2] https://github.com/MoinSebi/gfa2bin

`Nearest` - Position on a reference path

Linking nodes to the closest node of a reference path. Requested node must be part of any other path. A reference node is the closest node which can be found on any given reference path. The result does additionally return the reference position of this node.

Other commands

We also offer a range of commands to modify PLINK files. This can be done either by filtering based on allele frequencies and missing-rates, or remove certain genotypes and samples entirely. To control the genotypes, users can also explore the plain-text vcf-like file using `gfa2bin view`, or split and merge multiple files for better performance.

4.3 Materials and Methods

4.3.1 Datasets

The datasets were sourced from published resources. *Arabidopsis thaliana* assemblies for graph construction were obtained from the 1001 Genomes Plus project [58]. Short reads were from the 1001 Genomes Project [3] and corresponding phenotypes from multiple publications, collected and curated as previously published by Voichek and Weigel [137].

4.3.2 Kinship matrix

To facilitate the comparison between SNP-, k-mer and graph node-based GWAS, we used the same kinship matrix for all three. The kinship matrix was originally calculated on SNP data using the method from EMMA [66, 137].

4.3.3 Genome graph building

The *A. thaliana* graph was constructed with pggb [39]. The pggb workflow was executed with wfmash (v0.10.2-2-gb310bd1), seqwish (v0.7.8-3-gd9e7ab5), odgi (v0.8.2-92-gbfae0b3), and smoothxg (v0.6.8-31-g06bbf35). Our parameter set was -s 10000 -k 79 -n 27 -p 90 -P asm10.

4.3.4 Converting nodes to reference positions

We anchored all nodes tested in GWAS using `gfa2bin nearest` to specific locations on the TAIR10 reference genome. Specifically, we identified the shortest path (in base pairs) to any reference node and recorded the distance in base pairs. To denote the proximity to the respective anchor node in the reference, different markers or colors can be used in Manhattan plots (Figure 4.8).

4.3.5 Visualization

Manhattan plots were generated using the Python matplotlib library. To enhance clarity, we excluded all hits falling above a threshold of $\log_{10}(-2)$. Graphs were represented in two dimensions ("2D") using Bandage NG [142, 71]. The graph layouts show the significant node and its 50 neighboring nodes. For a linear ("1D") representation of the FLC region in the graph, `odgi viz` was employed. We extracted a subgraph using Bandage NG and manually added the represented path to facilitate detailed visualization.

4.3.6 Effect size

For effect size computation we used Point-Biserial Correlation. We selected this correlation technique to highlight extent and direction (positive or negative) of the relationship and get insights into whether two groups differ in terms of their average outcome.

4.3.7 GWAS validation

To compare the results from different GWAS inputs, we followed a published workflow [137]. This includes ranking nodes by their initial scoring, then using the 10,000 top ranked genotypes to calculate the exact *P* values, for which a permutation-based threshold is calculated. This process is repeated until all hits are in the top half of the initial ranking, passing more entries with each repetition.
SNP and k-mer hits for the set of phenotypes examined have been published [137]. Among the shared associations, we filtered out traits with fewer than 40 phenotyped accessions, or where there was a difference between SNP GWAS and a uniform distribution (Kolmogorow-Smirnow test, $p \leq 0.05$; adapted from [137]).

4.4 Results

In our *A. thaliana* case study, we first aligned 1,008 short-read datasets [3] to a genome variation graph containing 28 genomes [58]. We included 1,695 traits with multiple testing correction using hundredfold permutation for each trait. With the graph nodes, we found significant associations for 637 traits, compared to 668 traits with k-mers, and 594 traits with SNPs. The largest class comprised the 398 traits that had significant associations with all three inputs (Figure 4.4A). The largest overlap between exactly two inputs was observed for k-mers and graph nodes, 129, while SNPs had the highest number of unique hits, 88.

4.4.1 GWAS validation

In a direct comparison to SNPs, the use of graph nodes as input identified a greater number of significant hits in shared associations, along with a higher incidence of hits unique to the graph methodology (Figure 4.4C). Additionally, these graph node-based hits had on average lower *P* values, particularly when focusing on the top hit for each trait (Figures 4.4B and 4.4D). The high correlation and the positive long-tailed distribution of the ratios between *P* values from graph nodes and SNPs (see Figures 4.4B and 4.4D) validated our approach and the robustness of the underlying graph mapping for GWAS. In traits common to both inputs, we did not detect any exceptional outliers.

4.4.2 Flowering time

For a more in-depth validation, we used flowering time, which is a thoroughly investigated trait in *A. thaliana* [19]. Our graph node approach successfully re-identified several known associations [3], including with the loci *FLOWERING LOCUS T (FT), FLOWERING LOCUS C (FLC), DELAY OF GERMINATION 1 (DOG1)* and *VERNALIZATION INSENSITIVE 3 (VIN3)* (Figure 4.4A). Direct visualization and re-annotation revealed an insertion in *FLC*, depicted as a loop in the graph representation of genome sequences, as being responsible for the detected association with flowering time in our dataset (Figure 4.4B). The linear representation of genome sequences in Figure 4.4C confirmed the indel nature of the polymorphism, which is absent from the majority of our 28 reference genomes, including the original TAIR10 reference genome. The paths traversing the loop, which correspond to four of our 28 reference genomes in Figure 4.4C, are associated with later flowering (Table 4.1).

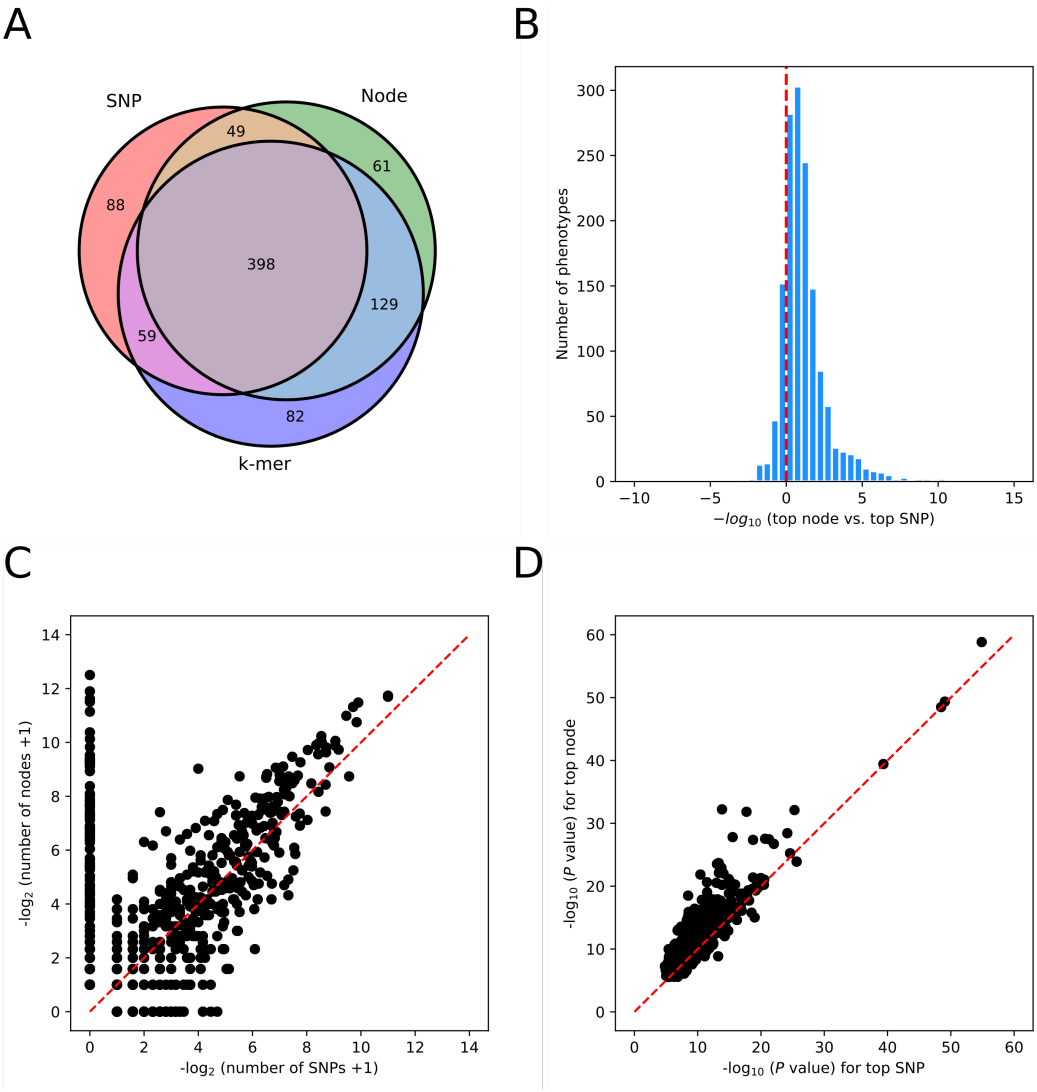


Fig. 4.4 **Comparison of SNP-, k-mer-, and graph node-based GWAS on 1,695 *A. thaliana* phenotypes.**
(A) Overlap between phenotypes with SNP, k-mer and graph node hits. **(B)** Ratio between top P values (expressed as $-\log_{10}$) for graph node and SNP hits. **(C)** Correlation of numbers of significant graph nodes versus SNPs. **(D)** Correlation of P values of top nodes with SNPs ($r = 0.91$).

4.4.3 Effect size

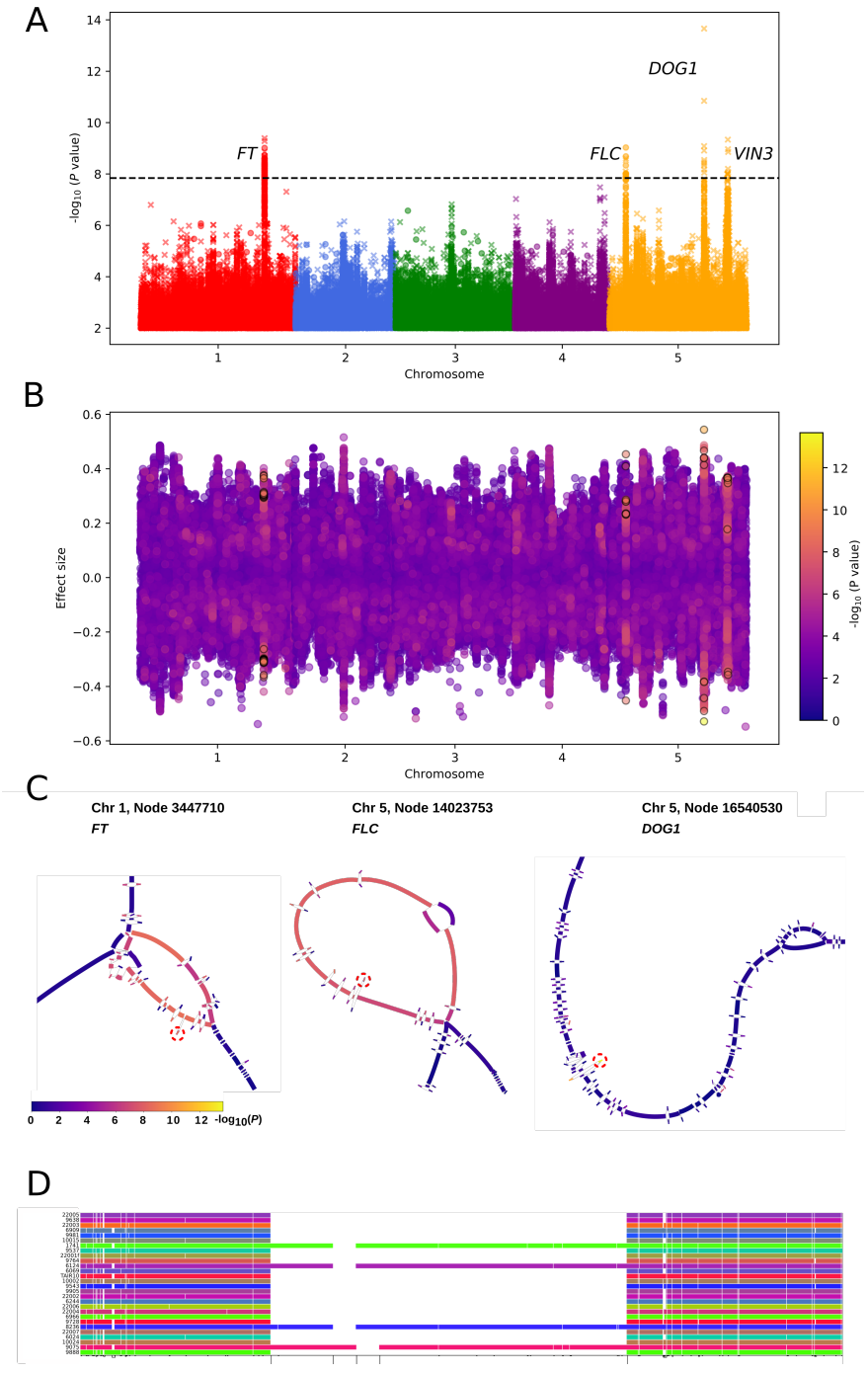
We calculated the effect size of presence/absence of nodes on flowering time at 10°C *(FT10)* without kinship correction. While the significant GWAS hits did not necessarily have the largest effect sizes – indicating the importance of kinship correction (Figure 4.4B), it was noteworthy that significant nodes at a locus included ones with opposite effects, indicating

that our approach identifies what appear to be "alternative" alleles. A similar bubble-like structure was also observed for the *FT* locus (Figure 4.4C). Even though the nodes along the whole structure were associated with elevated *P* values, the most significant graph node was found in one of the alleles with a size of 1 bp, indicating a short indel polymorphism. The *DOG1* locus showed a linear structure, with the most significant node representing a single SNP, comparable to traditional SNP-based GWAS (Figure 4.4D).

4.4.4 New associations

We visualized subgraphs for exemplary types of associations detected with graph nodes but not with SNPs (supplementary loci table available online[3]), to highlight the range of possible explanations. One example represents what is likely one of the most common causes, a complex region of the genome, with multiple loops in the graph, which would interfere with accurate mapping of short reads (Figure 4.6A). In another example, the causal variant was an indel, with two SNP alternative alleles. Interestingly, while not detectable with short-read based SNPs, this association was detectable with markers called from array hybridization [37], (Figure 4.6B). A third example represents likely another common case, where the association is with a SNP, but in a TE-rich complex region (Figure 4.6C). In agreement, there is a statistically significant difference in the number of TE hits between SNP-only and graph node-only hits (Figure 4.5).

[3]http://ftp.tuebingen.mpg.de/ebio/ibezrukov2/vorbrugg2025/GFA2BIN.supplemenatry_table.nodes_snps.xlsx



(Caption next page.)

Fig. 4.4 (Previous page.) **Interpretation of node-based GWAS.**
(A) Associations of graph nodes with flowering time at 10°C *(FT10)* as trait, shown on a linear representation after conversion to the TAIR10 standard reference. Nodes present in the reference are denoted by X. Nodes that are absent from the reference are denoted by "o". They were linked to the reference by finding the closest reference node, as described in the Methods. **(B)** Effect size of each node (linked to TAIR10 position) computed by point-biserial correlation and without kinship correction. Significant hits are highlighted with black outlines. **(C)** Subgraph visualization for associations with *FT* (4912 bp shown), *FLC* (2355 bp shown) and *DOG1* (1336 bp shown). Note the loop structure at *FLC*, which identifies sequences absent from the TAIR10 reference. **(D)** Representation of the *FLC* subgraph on a linear (graph node) scale.

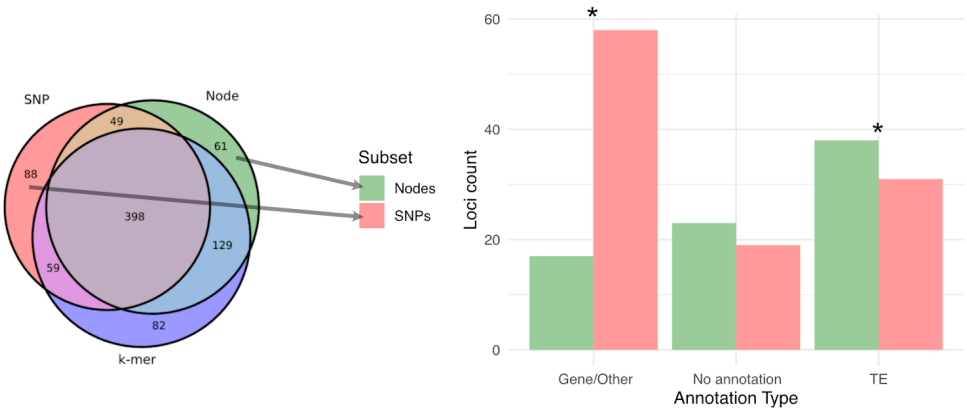


Fig. 4.5 **Differences in loci count for SNP-only and node-only associations based on annotation type.**
Each locus was counted only once. Differences for TE as well as for Gene/Other are statistically significant (asterisk, Fisher's exact test). Computed from supplementary loci table, with SNP locus 3 filtered out due to an excessive number of hits in a pericentromeric region.

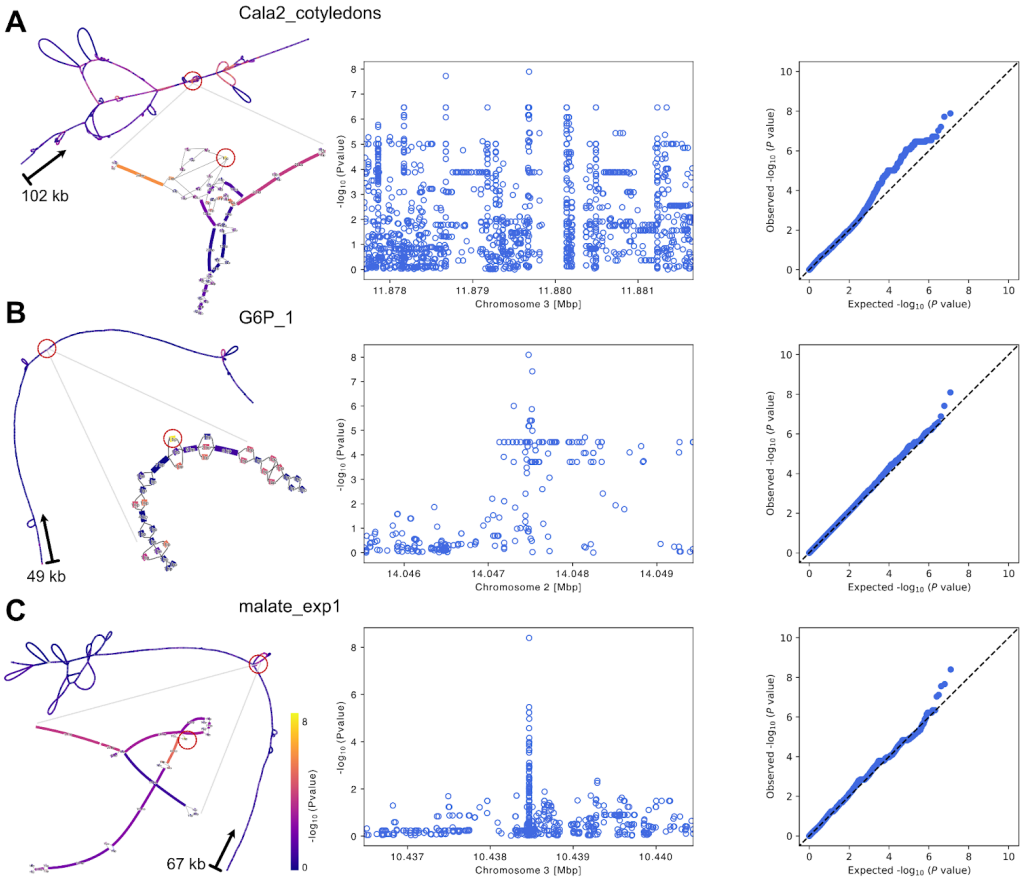


Fig. 4.6 **Subgraph visualization for selected associations where phenotypes were detected with node-based but not with SNP GWAS. (A)** An association in direct neighborhood to a gene in TAIR10, with a complex structure containing multiple loops. Ten of the 28 long-read genomes have the node, including 6966 (Sq-1), which well supported sporulation of *Hyaloperonospora arabidopsidis* Cala-23. Col-0 (TAIR10) lacks the node and did not support sporulation. **(B)** An association in a direct neighborhood to a gene in TAIR10. While this hit was not detected with SNPs called from short reads, it was detectable with SNP array data. **(C)** An association close to a transposable element in a very complex region with multiple loop structures. While the hit lies on a 1-bp long node, the surrounding region contains nodes almost 1 kb long.

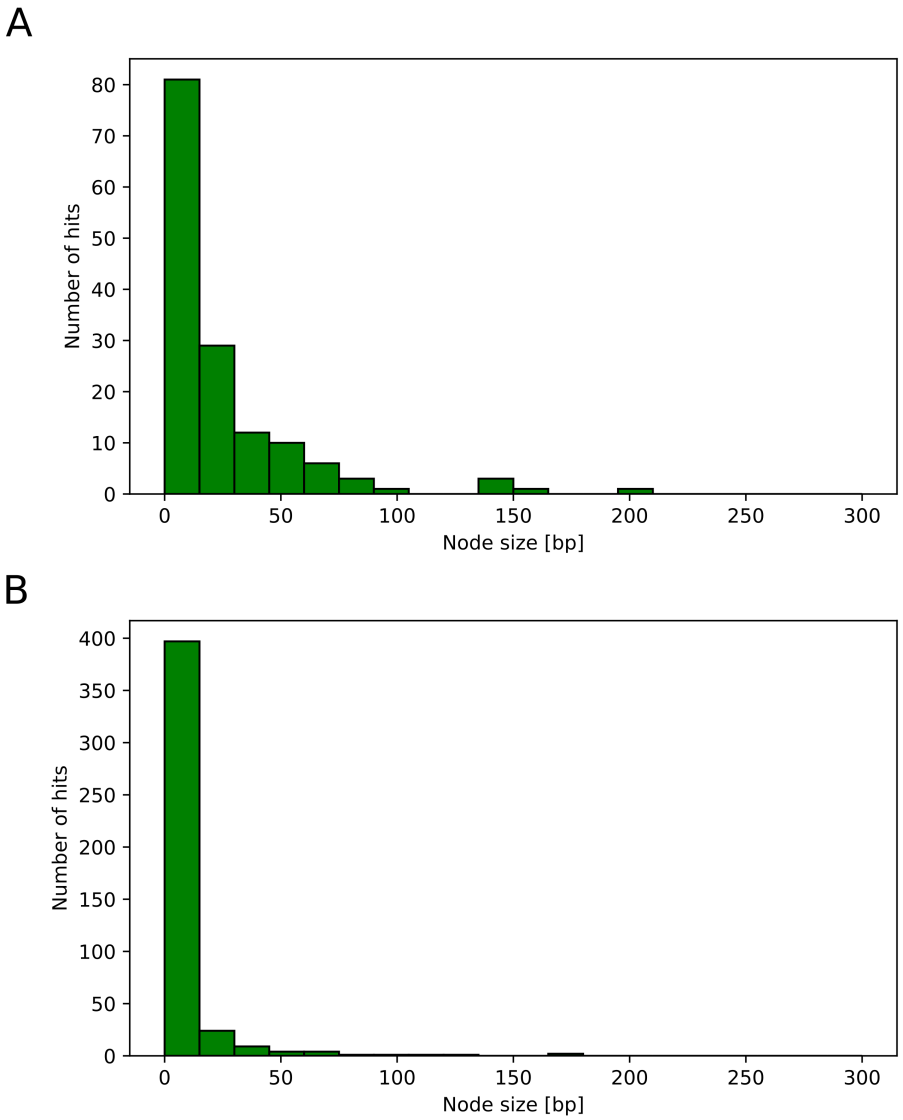


Fig. 4.7 **Node sizes of the top hits.**
(A) SNP-exclusive GWAS hits, transferred to the graph-coordinate system (nodes). **(B)** Shared SNP- and graph node-based hits. SNP-specific hits are mostly found in longer nodes with a median of 11 bp compared to 1 bp in the shared dataset.

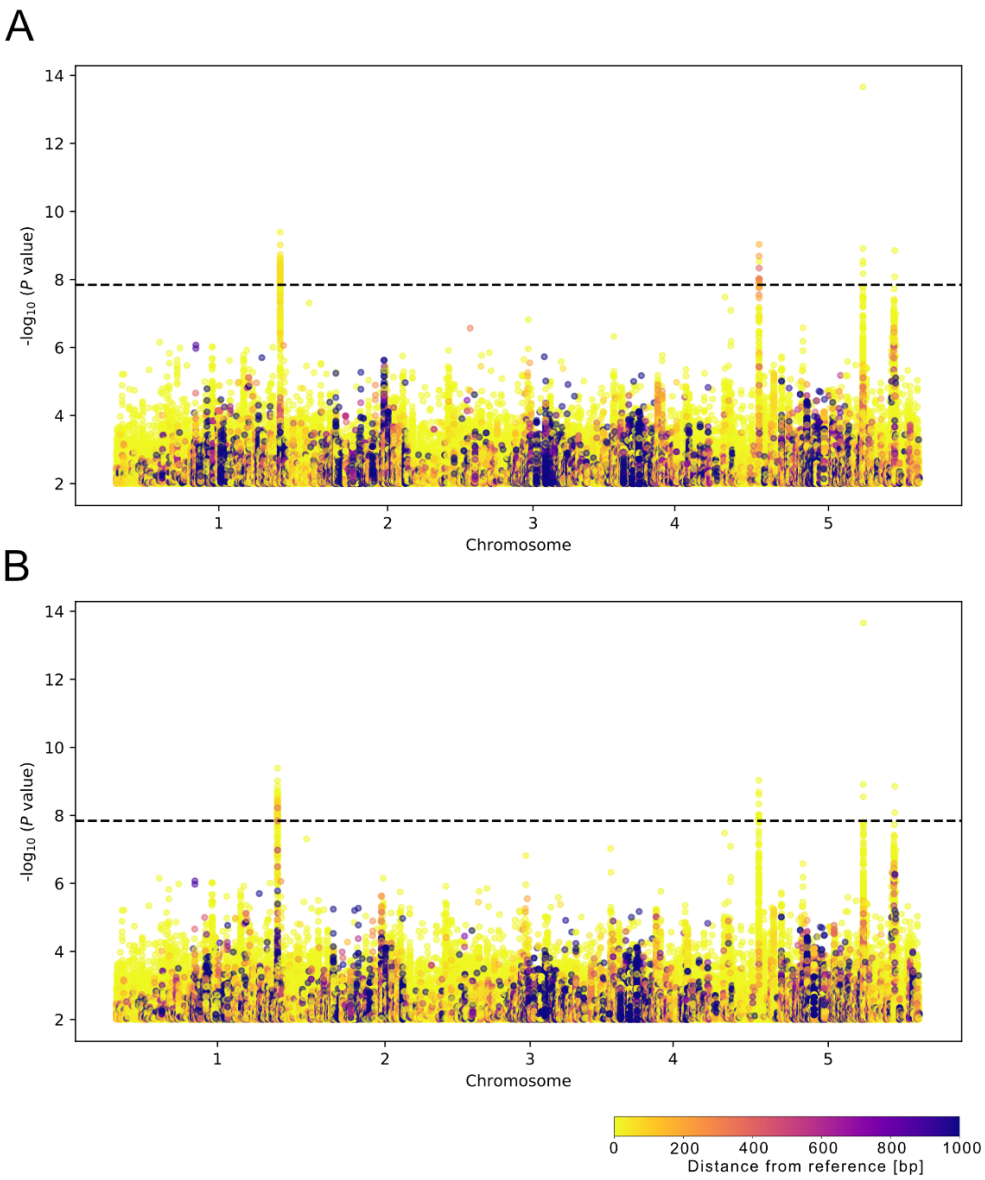


Fig. 4.8 **Manhattan plots for *FT10* trait projected onto different reference genomes.** Graph nodes were mapped to **(A)** TAIR10 or **(B)** KBS-Mac-74 (accession ID 1741). Overall results are similar, underscoring the robustness of the approach.

4.5 Discussion

We have shown the `gfa2bin` tool to be an effective and versatile solution for converting variation graphs to GWAS-ready formats. Starting from alignment of short reads to a graph that captures most common structural variants in the global *A. thaliana* population, we reproduced many associations that were identified before using conventional GWAS approaches with a single linear reference genome [3]. Going beyond SNP-based GWAS, graph-based GWAS identified more associations with greater statistical significance, but there were also associations that were identified only with SNPs or only with k-mers. This indicates that the improved representation of genetic diversity through variation graphs enhances the scope and effectiveness of GWAS, but that the best results are obtained by combining multiple methods. Our workflow allows for straightforward interpretation of graph node hits in the context of a standard reference genome, which in turn facilitate integration of results with those from SNP-based approaches. A drawback of k-mers is that k-mers not present in the standard reference genome can be difficult to interpret [137]. In principle, it should be possible to link k-mer hits to graph node-based hits by mapping k-mers to the genome graph.

Associations found only with SNPs can at least in some cases be traced back to graph nodes containing long sequences (Figure 4.7). We interpret this as the causal SNP not being captured in our set of 28 genomes used to construct the graph. If a single SNP is found in a longer node, this would only marginally reduce coverage of the node by the sampled short reads, and we would score the node as present regardless of the SNP. Conversely, multiple alternative alleles nor represented at a node in the graph, might all be scored as zero coverage. This would reduce power in association testing, if only some of these alternative alleles lead to phenotypic differences.
Our tool supports genotyping at base pair level, which, while computationally intensive, would aid in uncovering more associations and mitigating the effects of averaging factors. In addition, a larger number of diverse genomes to build the initial graph for short read mapping should provide for a greater set of nodes that can be tested for association with different phenotypes.

We not only provided a framework for coverage and node-based GWAS in variation graphs, but also support the usage of complex pggb graphs, which have not yet been used for mapping-based GWAS approaches. These graphs are designed to incorporate alignments of all input sequences and do not rely on pre-computed variation.
In the long term, to fully exploit all of genomic complexity, a more sophisticated representation of variants in graphs might be necessary. This would include the ability to represent and

identify both simple and complex structures, or combining multiple nodes into, e.g., bubbles with different alleles. Additionally, at the alignment level, a more nuanced representation of potential alternative alleles (variants), which may or may not be artifacts of the alignments, could enhance the identification of associations. It is crucial to recognize that, as in traditional SNP-based GWAS workflows, input quality is a primary determinant of ability to detect associations. As with SNP-based workflows, coverage depths and error rates in the short read data set used to call variants on the graph are one such determinant. In addition, the quality of the variation graph used for mapping will significantly influence the GWAS results. On the one hand, quality is determined by the number and accuracy of complete input genomes for building the graph. On the other hand, choices made when building the graph must be carefully considered, because the optimal graph will depend on the extent of variation present in the input genomes [39].

In future, instead of mapping short reads to a graph, one could use completely assembled genomes to genotype the paths in the graph. This would not only obviate the need for arbitrary thresholds to call node presence and absence, but would also allow for the use of copy number variants, inferred from the number of times a specific genome path visits the same node, in GWAS. In anticipation of this bright future, we have already implemented commands that could accommodate such analyses.

Table 4.1 **Geographic and flowering time of accessions used for graph construction.** This table presents the geographic coordinates (longitude and latitude) and flowering time data for the *A. thaliana* accessions used to build the graph [58]. Orange highlights accessions with the FLC insertion shown in Figure 4.4

Accession ID	Flowering time	Longitude	Latitude	Country
1741	-	-85.398	42.405	US
6024	106.75	13.37	55.75	Sweden
6069	107.25	18.28	62.95	Sweden
6124	129.50	13.31	55.84	Sweden
6244	90.75	18.47	62.92	Sweden
6909	70.50	-92.30	38.30	US
6966	62.25	-0.64	51.41	United Kingdom
8236	74.00	15.76	49.33	Czech Republic
9075	93.25	48.61	38.74	Azerbaijan
9537	64.25	-6.66	38.07	Spain
9543	94.33	-5.39	36.77	Spain
9638	88.25	80.86	51.73	Russia
9728	77.50	18.90	48.46	Slovakia
9764	69.75	35.84	34.10	Lebanon
9888	86.25	-3.31	40.93	Spain
9905	89.75	-4.01	40.76	Spain
9981	69.25	16.24	38.76	Italy
10002	61.00	9.04	48.53	Germany
10015	69.75	71.30	37.29	Afghanistan
10024	-	36.21	-2.87	Tanzania
22001	-	115.06	32.14	China
22002	-	108.61	27.94	China
22003	-	-4.10	34.09	Morocco
22004	-	-7.41	31.47	Morocco
22005	-	-17.13	32.75	Portugal
22006	-	-16.93	32.74	Portugal
22007	-	38.06	13.24	Ethiopia

Chapter 5

Conclusion & Outlook

5.1 Conclusion

Graphs have emerged as fundamental data structures in bioinformatics over the past several decades [112]. Their utility spans a wide range of applications, depending on the structure of the graph, the underlying data, and the algorithms applied. Despite these differences, their core objective is often similar: to represent entities with high similarity or direct relationships as connected nodes, while dissimilar entities are typically positioned further apart, often lacking direct connections.

One of the earliest uses of graph-based representations in biology was in modeling protein-protein interaction networks and metabolic pathways [36, 10]. As biological data volumes grew, especially with the advent of high-throughput sequencing, graphs became increasingly important for tasks such as genome assembly, where structures like De Bruijn graphs (DBGs) and overlap-consensus graphs are used to reconstruct genomes from short or long reads by identifying a "golden path" through overlapping sequences [98, 112].

When reference genomes are available, graphs can be used to represent splicing variability (e.g., splice graphs) and structural or small-scale variation across individuals (e.g., variation graphs) [41]. These allow a more comprehensive view of population-level diversity and support the construction of graph-based references, a major shift from traditional linear genome models.

In recent years, pangenome graphs have gained prominence, particularly in projects such as the Human Pangenome Reference Consortium (HPRC) [81]. These graphs aim to unify the genomic diversity of multiple individuals into a single, non-linear reference structure, improving accuracy in downstream analyses such as variant calling and genotyping.

Tools like `VG`, `ODGI`, and `minigraph` have been developed to construct and manipulate genome graphs, with data formats such as GFA (Graphical Fragment Assembly) and GAF (Graph Alignment Format) serving as emerging standards for interoperability [41, 46, 53].

Despite their promise, graph-based approaches come with challenges. They remain computationally intensive, require specialized algorithms, and pose difficulties in visualization, storage, and interpretability, especially for non-expert users. Addressing these issues will be critical for broadening the adoption of genome graphs and realizing their full potential in genomics research.

The methods and findings presented in this thesis build upon this context, investigating both the technical challenges and practical opportunities of genome graph analysis through a series of novel algorithms and applied case studies.

This thesis explores the structure and interpretability of genome graphs through the development of novel tools and methods. It focused on how graph topology is influenced by parameter and method choices, the extraction and interpretation of structural metrics, and the application of graph-based GWAS, leveraging internal variation or using the graph as variation-aware reference for variation calling.

`Gretl` was developed as a statistical toolkit to evaluate and compare genome graphs in a standardized, scalable way. It computes a wide range of graph- and path-based metrics, allowing for structured analysis across different graphs, parameter settings, and construction methods. Within this thesis, `gretl` served as the core analytical method for multiple investigations.

A key application was the systematic exploration of parameter sensitivity in the `PGGB` workflow using yeast genomes. Here, `gretl` revealed that graph structure is highly sensitive to the segment length ('s') and alignment threshold ('p') parameters. These two settings were found to influence fundamental features such as node size, compression, and overall complexity. Other parameters like 'n', 'k', and 'asm' showed less consistent effects in our dataset. However, even when individual parameters appear to have limited impact, the right combination of settings can lead to substantial changes in graph topology, especially when aiming to generate extreme or edge-case graph structures. This demonstrated that even subtle parameter changes, when combined, can significantly reshape the resulting graph, particularly in repetitive or compact genomes.

In addition to parameter testing, `gretl` enabled a detailed comparison between graphs built using `PGGB` and `Minigraph-Cactus` (MC) [39, 53]. Despite using the same input sequences, the two tools produced markedly different graphs, with PGGB retaining more local variation and structural complexity, while MC generated more linear and conservative

graphs by omitting highly divergent regions. Metrics derived by `gretl`, such as similarity, node length, and edge inversion, made these differences both quantifiable and interpretable.

Notably, `gretl` also supports the integration of graph annotations, allowing users to link structural metrics with external features such as gene models or functional regions (Appendix B.1 for examples). This makes it possible to explore the biological relevance of structural differences in a more targeted and informative manner.

Overall, `gretl` provided the statistical foundation for identifying how both parameter choices and construction algorithms affect genome graph topology. Its ability to reduce complex structures into interpretable summaries makes it a valuable tool for graph-based genome research and an important step toward reproducible, comparative, and biologically meaningful graph analysis.

The analysis of the 1001G+ *Arabidopsis thaliana* graph revealed clear population structure, with two major clades corresponding to European/Asian and African/Madeiran/relict accessions. These clusters were consistent across multiple approaches, including Mash distance, PCA, and structural variant detection, and SNPs, reinforcing the robustness of the observed patterns. The presence of substructure, especially among relict and Madeiran accessions, highlights the strength of the selected dataset. The inclusion of newly added genomes provides a clearer and more detailed picture of the historically ambiguous relict groups, helping to resolve their phylogeographic value and evolutionary distinctiveness.

We showed that the core pangenome constitutes approximately 60% of the total genome size, with most of the remaining sequence also shared across multiple accessions. Nevertheless, less-shared regions were primarily located on chromosomes, highlighting technical limitations such as the influence of repetitive elements and under-resolved centromeric regions in CLR-based assemblies. A closer analysis of pangenome distribution and growth revealed that most newly added sequences are private and tend to occur in intergenic regions. The lack of saturation suggests that each newly added genome contributes unique sequences, with more genetically distinct accessions contributing disproportionately more to the expanding pangenome.

Bubble statistics provided a finer-grained view of variation, revealing a landscape dominated by small, biallelic bubbles but punctuated by larger, complex and often nested structures. Many sample-specific bubbles likely reflect real duplications or TE insertions, though some may result from normalization artifacts or alignment ambiguity. The exponential decline in bubble size and U-shaped traversal distribution underscore the coexistence of shared and highly individual variants within the population.

Finally, a comparison with Pannagram illustrated the challenges of distinguishing true variation from alignment- or repeat-induced artifacts. While PGGB tends to represent divergence as long SVs, Pannagram fragments these into clusters of short events. The divergence in results between these tools reflects underlying algorithmic differences, reinforcing the need to align methods with biological questions rather than seeking a single "correct" output.

Overall, the *Arabidopsis* graph analysis demonstrates the power of genome graphs to reveal both broad and fine-scale population structure, while also exposing the methodological challenges of interpreting complex variation.

Building on graph-derived statistics, we explored graph-based genome-wide association studies (GWAS) using the `gfa2bin` tool. By mapping short reads to a variation graph constructed from 28 *A. thaliana* genomes, we were able to reproduce known SNP-based associations and identify additional signals with greater statistical significance. This demonstrates the enhanced sensitivity of graph-based GWAS for capturing structural and presence/absence variation that is often missed in linear reference-based approaches.

In our analysis, we also identified a promising novel association, though its biological relevance remains to be confirmed. Importantly, the power and resolution of graph-based GWAS are expected to increase as more genomes are incorporated into the graph. A more diverse and complete graph reduces reference bias by incorporating a broader range of large variants that are not missing in linear references, ensuring they are directly represented rather than indirectly inferred or missed entirely.

Our method supports mapping-based genotyping on PGGB graphs and allows direct integration with standard reference coordinates, facilitating downstream interpretation. However, some limitations remain. Long nodes containing uncaptured SNPs can reduce association power, and variants absent from the input graph but present in newly aligned samples are not represented. This restricts the ability to detect alternative alleles not encoded in the original graph.

While edge-based GWAS is a potential extension, it sacrifices base-pair resolution and operates only at the level of node connectivity, preventing direct base-level coverage analysis. Due to time constraints, we were unable to implement this in the current study. Similarly, approaches that call variants directly (vg call) from the graph and produce VCFs may offer better support for allele-specific analyses, but they are more complex and computationally intensive.

Despite these challenges, our method offers a practical and scalable solution, particularly suited for large cohorts, demonstrated here on over 800 samples. Its relative simplicity, combined with the ability to capture structural variation, makes it a compelling alternative

to traditional pipelines. Ultimately, the goal is to perform GWAS directly within the graph itself, without relying on external variant calling or re-sequencing. However, this requires graphs that are sufficiently complete: (1) capturing the majority of relevant variation, and (2) providing adequate sample representation for each variant to support meaningful association testing. Until these conditions are met, re-sequencing-based approaches, whether using `gfa2bin` or graph-derived variant calls, will continue to serve as necessary intermediates toward fully graph-native GWAS.

5.2 Bubble detection – Limitations of current graph-based approaches and strategies for overcoming them

The accurate detection and interpretation of structural variation in genome graphs is a central challenge in pangenome research [110, 52]. As genome graphs become increasingly used to represent genetic diversity across populations, the need for methods that can extract meaningful biological variation directly from the graph becomes more urgent [41, 77]. Traditional bubble detection approaches, such as those implemented in the VG toolkit or BubbleGun, are often limited by strict assumptions, such as acyclicity or lack of directionality, that do not hold in complex, real-world graphs [105, 23]. Furthermore, most existing tools offer limited insight into how variation relates to individual samples or paths in the graph [109].

Motivated by these limitations, this work investigated a new approach to bubble detection that leverages bifurcation patterns between paths. Rather than relying solely on topological definitions of bubbles, this method identifies regions where paths diverge and later reconverge, capturing variation in a way that is both intuitive and biologically grounded. The key idea is that a bubble begins when two or more paths split from a common node and ends when they meet again, making bifurcation a natural indicator of genomic variation (Appendix B.3).

To implement this concept, the algorithm systematically compares all pairs of paths and identifies shared nodes where such bifurcations occur (Appendix B.3.3). A central feature of the method is its ability to detect nested bubbles, cases where smaller variants are contained within larger, structurally more complex regions, thus offering a hierarchical view of genome variation. In addition, the method supports output formats that can be easily used for downstream analyses, such as BED-based interval annotation and traversal statistics.

The advantages of this approach are several: it is path-aware, scalable through parallelization, and applicable to a wide range of graph structures, including those with cycles and complex repeat regions. It also produces outputs that are easy to interpret in the context of sample-level variation.

Nonetheless, the algorithm also faces bottlenecks. The all-versus-all comparison of paths introduces quadratic scaling, which can become computationally intensive in large graphs. A more serious challenge lies in handling highly repetitive regions, such as centromeres, where repeated nodes and complex edge patterns create vast amounts of redundant data. While technical improvements, such as replacing hash-based data structures with compact vectors, have significantly improved performance, further optimization is needed to reduce memory usage and runtime.

In summary, this bifurcation-based bubble detection strategy offers a biologically motivated and practically useful alternative to existing approaches, enabling a more nuanced understanding of structural variation in genome graphs. The current proof-of-concept implementation and algorithm may lay the groundwork for future tools that integrate graph topology with sample-specific variation in a scalable and interpretable manner.

The BVD bubble detection algorithm developed in this thesis introduces a novel, path-aware approach to identifying variation structures in genome graphs. In contrast to traditional superbubble definitions based on acyclic subgraphs, our method leverages bifurcation events, defined by divergence and reconvergence of paths, to detect candidate variation regions. This path-centric design allows the algorithm to operate directly on pairwise path comparisons, capturing fine-scale variation missed by global, topology-only methods.

While highly parallelizable, the approach faces challenges in terms of scalability, particularly due to its quadratic complexity with respect to the number of paths. Highly repetitive nodes, especially in centromeric regions, lead to large indexes and inflated output, posing both performance and interpretability issues. Nevertheless, the algorithm handles a wide range of variation types, from simple SNPs to complex nested structures, and supports downstream analysis through additional features like nestedness detection.

Future improvements could include smarter path sampling, filtering of repetitive regions, and graph-aware optimizations. As the method is still under development, current evaluations are based on selected examples, which have helped identify key bottlenecks and guide further refinement. Despite these limitations, BVD lays a strong foundation for flexible and biologically meaningful variation detection in genome graphs. A public open-source release is planned to support broader adoption and continued development.

5.3 Outlook

As genome graphs continue to gain traction in genomics, several key challenges and future directions have become clear [77]. Rather than choosing between two competing strategies, expanding pangenome graphs to incorporate increasing numbers of genomes, or maintaining high-quality reference graphs for ongoing resequencing, both approaches can coexist, as they serve distinct purposes. However, each strategy requires fundamentally different graph structures: large, inclusive pangenomes demand highly scalable, variation-rich representations, while stable reference graphs prioritize consistency, interpretability, and efficient mapping. In both cases, improved methods for graph alignment are essential, as current tools still struggle with accuracy in complex or repetitive regions [41, 116].

Coverage-based analysis and variation calling on genome graphs offer a promising alternative to traditional SNP-based approaches, but current pipelines are still under development and show variability in performance, particularly in complex or repetitive regions [46]. While these methods can efficiently leverage mapping data to detect presence/absence and structural variation, they often fall short in capturing finer-scale variation with high confidence.

For now, resequencing-based strategies remain the more robust and widely applicable solution. This is largely due to the limited availability of chromosome-scale assemblies needed to construct comprehensive, fully representative pangenome graphs. In this context, accurate short-read alignment and reliable post-processing steps remain critical. As assembly quality and graph frameworks improve, coverage-based methods are likely to become more powerful, but for the time being, resequencing continues to provide a necessary backbone for graph-based analysis.

One major limitation in current graph frameworks is the lack of robust support for adding new sequences or genomes to existing graphs. Most tools are designed for static, one-time construction, requiring complete reassembly when new data is introduced. This limitation poses a significant barrier for population-scale studies, where continuous sampling and integration of new genomes are common.

Early versions of the VG toolkit included preliminary functionality for graph augmentation based on called variants, allowing incremental updates without full reconstruction [41]. However, this approach appears to be outdated and is no longer actively supported. While it may work well for integrating small, localized variation, adding larger structural variants remains a major challenge. Placement of such variants requires contextual understanding of the surrounding graph topology, and naive insertion may disrupt existing paths or misrepresent variant relationships.

Moreover, sequential integration of genomes introduces bias, as the graph structure becomes dependent on the order in which samples are added [53]. A more ideal solution

would involve an all-vs-all comparison framework for graph extension, which ensures that new variation is incorporated symmetrically and without topological distortion. However, implementing such a system poses significant technical challenges and would require new data structures and alignment strategies. Nonetheless, the ability to modularly update graphs remains a critical need for making genome graphs scalable and usable in real-time population genomics.

A potential strategy to improve both scalability and interpretability is to construct smaller, functionally annotated subgraphs, such as gene- or locus-specific regions, which can later be concatenated into a unified graph. Using genome annotation to guide this segmentation would reduce complexity and focus computation on biologically relevant regions. In particular, prioritizing low-complexity regions can significantly simplify graph construction and improve the clarity of downstream analyses.

This modular approach also opens the door to using different graph construction parameters for different regions of the genome. For instance, highly repetitive or structurally complex loci could be processed with more stringent settings, while more conserved regions may benefit from relaxed thresholds to retain fine-scale variation. To enable such workflows, preliminary steps, such as scanning input genomes or building a lightweight "dummy" graph, could be used to preidentify regions of high complexity. These insights would then inform region-specific parameter choices during the main graph construction process, offering a flexible and adaptive pipeline tailored to genomic architecture.

Altogether, addressing these limitations and scaling strategies will be essential to fully realize the potential of genome graphs as reference structures for comparative and population genomics.

In summary, this thesis contributes both tools and insights that advance the interpretation and application of genome graphs. From statistical profiling with `gretl`, to graph-based GWAS with `gfa2bin`, and novel approaches to variation detection, each component addresses a key challenge in making genome graphs more accessible, analyzable, and biologically informative. By applying these tools to real datasets like the 1001G+ project, we demonstrate that graph-based methods can provide meaningful and sometimes unique perspectives on genome variation and structure.

At the same time, this work highlights current limitations, particularly in scaling, variant extraction, and graph extension, and the need for continued development of methods that bridge algorithmic efficiency with biological relevance. The tools and ideas presented here are intended to be made available to the community and may serve as building blocks for future improvements in graph-based genomics.

As genome graphs evolve beyond prototypes into widely adopted frameworks, the integration of flexible, modular tools will be critical. The contributions presented here aim to support that transition, and we hope they will contribute to establishing graph-based methods as a practical and robust foundation for the next generation of genome research.

References

[1] (1999). DNA viruses: A practical approach.

[2] 1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., and Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, 526(7571):68–74.

[3] 1001 Genomes Consortium (2016). 1,135 genomes reveal the global pattern of polymorphism in arabidopsis thaliana. *Cell*, 166(2):481–491.

[4] 3,000 rice genomes project (2014). The 3,000 rice genomes project. *Gigascience*, 3(1):7.

[5] Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., George, R. A., Lewis, S. E., Richards, S., Ashburner, M., Henderson, S. N., Sutton, G. G., Wortman, J. R., Yandell, M. D., Zhang, Q., Chen, L. X., Brandon, R. C., Rogers, Y. H., Blazej, R. G., Champe, M., Pfeiffer, B. D., Wan, K. H., Doyle, C., Baxter, E. G., Helt, G., Nelson, C. R., Gabor, G. L., Abril, J. F., Agbayani, A., An, H. J., Andrews-Pfannkoch, C., Baldwin, D., Ballew, R. M., Basu, A., Baxendale, J., Bayraktaroglu, L., Beasley, E. M., Beeson, K. Y., Benos, P. V., Berman, B. P., Bhandari, D., Bolshakov, S., Borkova, D., Botchan, M. R., Bouck, J., Brokstein, P., Brottier, P., Burtis, K. C., Busam, D. A., Butler, H., Cadieu, E., Center, A., Chandra, I., Cherry, J. M., Cawley, S., Dahlke, C., Davenport, L. B., Davies, P., de Pablos, B., Delcher, A., Deng, Z., Mays, A. D., Dew, I., Dietz, S. M., Dodson, K., Doup, L. E., Downes, M., Dugan-Rocha, S., Dunkov, B. C., Dunn, P., Durbin, K. J., Evangelista, C. C., Ferraz, C., Ferriera, S., Fleischmann, W., Fosler, C., Gabrielian, A. E., Garg, N. S., Gelbart, W. M., Glasser, K., Glodek, A., Gong, F., Gorrell, J. H., Gu, Z., Guan, P., Harris, M., Harris, N. L., Harvey, D., Heiman, T. J., Hernandez, J. R., Houck, J., Hostin, D., Houston, K. A., Howland, T. J., Wei, M. H., Ibegwam, C., Jalali, M., Kalush, F., Karpen, G. H., Ke, Z., Kennison, J. A., Ketchum, K. A., Kimmel, B. E., Kodira, C. D., Kraft, C., Kravitz, S., Kulp, D., Lai, Z., Lasko, P., Lei, Y., Levitsky, A. A., Li, J., Li, Z., Liang, Y., Lin, X., Liu, X., Mattei, B., McIntosh, T. C., McLeod, M. P., McPherson, D., Merkulov, G., Milshina, N. V., Mobarry, C., Morris, J., Moshrefi, A., Mount, S. M., Moy, M., Murphy, B., Murphy, L., Muzny, D. M., Nelson, D. L., Nelson, D. R., Nelson, K. A., Nixon, K., Nusskern, D. R., Pacleb, J. M., Palazzolo, M., Pittman, G. S., Pan, S., Pollard, J., Puri, V., Reese, M. G., Reinert, K., Remington, K., Saunders, R. D., Scheeler, F., Shen, H., Shue, B. C., Sidén-Kiamos, I., Simpson, M., Skupski, M. P., Smith, T., Spier, E., Spradling, A. C., Stapleton, M., Strong, R., Sun, E., Svirskas, R., Tector, C., Turner, R., Venter, E., Wang, A. H., Wang, X., Wang, Z. Y., Wassarman, D. A., Weinstock, G. M., Weissenbach, J., Williams, S. M., WoodageT, Worley, K. C., Wu, D., Yang, S., Yao, Q. A., Ye, J., Yeh,

R. F., Zaveri, J. S., Zhan, M., Zhang, G., Zhao, Q., Zheng, L., Zheng, X. H., Zhong, F. N., Zhong, W., Zhou, X., Zhu, S., Zhu, X., Smith, H. O., Gibbs, R. A., Myers, E. W., Rubin, G. M., and Venter, J. C. (2000). The genome sequence of drosophila melanogaster. *Science*, 287(5461):2185–2195.

[6] Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant arabidopsis thaliana. *Nature*, 408(6814):796–815.

[7] Armstrong, J., Hickey, G., Diekhans, M., Fiddes, I. T., Novak, A. M., Deran, A., Fang, Q., Xie, D., Feng, S., Stiller, J., Genereux, D., Johnson, J., Marinescu, V. D., Alföldi, J., Harris, R. S., Lindblad-Toh, K., Haussler, D., Karlsson, E., Jarvis, E. D., Zhang, G., and Paten, B. (2020). Progressive cactus is a multiple-genome aligner for the thousand-genome era. *Nature*, 587(7833):246–251.

[8] Avery, O. T., MacLeod, C. M., and McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type iii. *Journal of Experimental Medicine*, 79(2):137–158.

[9] Ballouz, S., Dobin, A., and Gillis, J. A. (2019). Is it time to change the reference genome? *Genome Biol.*, 20(1):159.

[10] Bartel, P. L., Roecklein, J. A., SenGupta, D., and Fields, S. (1996). A protein linkage map of escherichia coli bacteriophage T7. *Nat. Genet.*, 12(1):72–77.

[11] Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., Irving, L. J., Karbelashvili, M. S., Kirk, S. M., Li, H., Liu, X., Maisinger, K. S., Murray, L. J., Obradovic, B., Ost, T., Parkinson, M. L., Pratt, M. R., Rasolonjatovo, I. M. J., Reed, M. T., Rigatti, R., Rodighiero, C., Ross, M. T., Sabot, A., Sankar, S. V., Scally, A., Schroth, G. P., Smith, M. E., Smith, V. P., Spiridou, A., Torrance, P. E., Tzonev, S. S., Vermaas, E. H., Walter, K., Wu, X., Zhang, L., Alam, M. D., Anastasi, C., Aniebo, I. C., Bailey, D. M. D., Bancarz, I. R., Banerjee, S., Barbour, S. G., Baybayan, P. A., Benoit, V. A., Benson, K. F., Bevis, C., Black, P. J., Boodhun, A., Brennan, J. S., Bridgham, J. A., Brown, R. C., Brown, A. A., Buermann, D. H., Bundu, A. A., Burrows, J. C., Carter, N. P., Castillo, N., Chiara E Catenazzi, M., Chang, S., Neil Cooley, R., Crake, N. R., Dada, O. O., Diakoumakos, K. D., Dominguez-Fernandez, B., Earnshaw, D. J., Egbujor, U. C., Elmore, D. W., Etchin, S. S., Ewan, M. R., Fedurco, M., Fraser, L. J., Fuentes Fajardo, K. V., Scott Furey, W., George, D., Gietzen, K. J., Goddard, C. P., Golda, G. S., Granieri, P. A., Green, D. E., Gustafson, D. L., Hansen, N. F., Harnish, K., Haudenschild, C. D., Heyer, N. I., Hims, M. M., Ho, J. T., Horgan, A. M., Hoschler, K., Hurwitz, S., Ivanov, D. V., Johnson, M. Q., James, T., Huw Jones, T. A., Kang, G.-D., Kerelska, T. H., Kersey, A. D., Khrebtukova, I., Kindwall, A. P., Kingsbury, Z., Kokko-Gonzales, P. I., Kumar, A., Laurent, M. A., Lawley, C. T., Lee, S. E., Lee, X., Liao, A. K., Loch, J. A., Lok, M., Luo, S., Mammen, R. M., Martin, J. W., McCauley, P. G., McNitt, P., Mehta, P., Moon, K. W., Mullens, J. W., Newington, T., Ning, Z., Ling Ng, B., Novo, S. M., O'Neill, M. J., Osborne, M. A., Osnowski, A., Ostadan, O., Paraschos, L. L., Pickering, L., Pike, A. C., Pike, A. C., Chris Pinkard, D., Pliskin, D. P., Podhasky, J., Quijano, V. J., Raczy, C., Rae, V. H., Rawlings, S. R., Chiva Rodriguez, A., Roe, P. M.,

Rogers, J., Rogert Bacigalupo, M. C., Romanov, N., Romieu, A., Roth, R. K., Rourke, N. J., Ruediger, S. T., Rusman, E., Sanches-Kuiper, R. M., Schenker, M. R., Seoane, J. M., Shaw, R. J., Shiver, M. K., Short, S. W., Sizto, N. L., Sluis, J. P., Smith, M. A., Ernest Sohna Sohna, J., Spence, E. J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C. L., Turcatti, G., Vandevondele, S., Verhovsky, Y., Virk, S. M., Wakelin, S., Walcott, G. C., Wang, J., Worsley, G. J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J. C., Hurles, M. E., McCooke, N. J., West, J. S., Oaks, F. L., Lundberg, P. L., Klenerman, D., Durbin, R., and Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59.

[12] Brandt, D. Y. C., Aguiar, V. R. C., Bitarello, B. D., Nunes, K., Goudet, J., and Meyer, D. (2015). Mapping bias overestimates reference allele frequencies at the HLA genes in the 1000 genomes project phase I data. *G3 (Bethesda)*, 5(5):931–941.

[13] Burke, D. T., Carle, G. F., and Olson, M. V. (1987). Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science*, 236(4803):806–812.

[14] Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S., Allen, N., Donnelly, P., and Marchini, J. (2018). The UK biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209.

[15] Chaisson, M. J. P., Sanders, A. D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E. J., Rodriguez, O. L., Guo, L., Collins, R. L., Fan, X., Wen, J., Handsaker, R. E., Fairley, S., Kronenberg, Z. N., Kong, X., Hormozdiari, F., Lee, D., Wenger, A. M., Hastie, A. R., Antaki, D., Anantharaman, T., Audano, P. A., Brand, H., Cantsilieris, S., Cao, H., Cerveira, E., Chen, C., Chen, X., Chin, C.-S., Chong, Z., Chuang, N. T., Lambert, C. C., Church, D. M., Clarke, L., Farrell, A., Flores, J., Galeev, T., Gorkin, D. U., Gujral, M., Guryev, V., Heaton, W. H., Korlach, J., Kumar, S., Kwon, J. Y., Lam, E. T., Lee, J. E., Lee, J., Lee, W.-P., Lee, S. P., Li, S., Marks, P., Viaud-Martinez, K., Meiers, S., Munson, K. M., Navarro, F. C. P., Nelson, B. J., Nodzak, C., Noor, A., Kyriazopoulou-Panagiotopoulou, S., Pang, A. W. C., Qiu, Y., Rosanio, G., Ryan, M., Stütz, A., Spierings, D. C. J., Ward, A., Welch, A. E., Xiao, M., Xu, W., Zhang, C., Zhu, Q., Zheng-Bradley, X., Lowy, E., Yakneen, S., McCarroll, S., Jun, G., Ding, L., Koh, C. L., Ren, B., Flicek, P., Chen, K., Gerstein, M. B., Kwok, P.-Y., Lansdorp, P. M., Marth, G. T., Sebat, J., Shi, X., Bashir, A., Ye, K., Devine, S. E., Talkowski, M. E., Mills, R. E., Marschall, T., Korbel, J. O., Eichler, E. E., and Lee, C. (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.*, 10(1):1784.

[16] Chen, S., Krusche, P., Dolzhenko, E., Sherman, R. M., Petrovski, R., Schlesinger, F., Kirsche, M., Bentley, D. R., Schatz, M. C., Sedlazeck, F. J., and Eberle, M. A. (2019). Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome Biol.*, 20(1):291.

[17] Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., and Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods*, 18(2):170–175.

[18] Chin, C.-S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., Cramer, G. R., Delledonne, M., Luo, C., Ecker, J. R., Cantu, D., Rank, D. R., and Schatz, M. C. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods*, 13(12):1050–1054.

[19] Cho, L.-H., Yoon, J., and An, G. (2017). The control of flowering time by environmental factors. *Plant J.*, 90(4):708–719.

[20] Clarke, J., Wu, H.-C., Jayasinghe, L., Patel, A., Reid, S., and Bayley, H. (2009). Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.*, 4(4):265–270.

[21] Cork, J. M. and Purugganan, M. D. (2005). High-diversity genes in the arabidopsis genome. *Genetics*, 170(4):1897–1911.

[22] Crick, F. H. (1958). On protein synthesis. *Symp. Soc. Exp. Biol.*, 12:138–163.

[23] Dabbaghie, F., Ebler, J., and Marschall, T. (2022). BubbleGun: enumerating bubbles and superbubbles in genome graphs. *Bioinformatics*, 38(17):4217–4219.

[24] Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., and 1000 Genomes Project Analysis Group (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158.

[25] Degner, J. F., Marioni, J. C., Pai, A. A., Pickrell, J. K., Nkadori, E., Gilad, Y., and Pritchard, J. K. (2009). Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, 25(24):3207–3212.

[26] Du, Z.-Z., He, J.-B., and Jiao, W.-B. (2024). A comprehensive benchmark of graph-based genetic variant genotyping algorithms on plant genomes for creating an accurate ensemble pipeline. *Genome Biol.*, 25(1):91.

[27] Durvasula, A., Fulgione, A., Gutaker, R. M., Alacakaptan, S. I., Flood, P. J., Neto, C., Tsuchimatsu, T., Burbano, H. A., Picó, F. X., Alonso-Blanco, C., and Hancock, A. M. (2017). African genomes illuminate the early history and transition to selfing in arabidopsis thaliana. *Proc. Natl. Acad. Sci. U. S. A.*, 114(20):5213–5218.

[28] Earl, D., Bradnam, K., St John, J., Darling, A., Lin, D., Fass, J., Yu, H. O. K., Buffalo, V., Zerbino, D. R., Diekhans, M., Nguyen, N., Ariyaratne, P. N., Sung, W.-K., Ning, Z., Haimel, M., Simpson, J. T., Fonseca, N. A., Birol, I., Docking, T. R., Ho, I. Y., Rokhsar, D. S., Chikhi, R., Lavenier, D., Chapuis, G., Naquin, D., Maillet, N., Schatz, M. C., Kelley, D. R., Phillippy, A. M., Koren, S., Yang, S.-P., Wu, W., Chou, W.-C., Srivastava, A., Shaw, T. I., Ruby, J. G., Skewes-Cox, P., Betegon, M., Dimon, M. T., Solovyev, V., Seledtsov, I., Kosarev, P., Vorobyev, D., Ramirez-Gonzalez, R., Leggett, R., MacLean, D., Xia, F., Luo, R., Li, Z., Xie, Y., Liu, B., Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F. J., Yin, S., Sharpe, T., Hall, G., Kersey, P. J., Durbin, R., Jackman, S. D., Chapman, J. A., Huang, X., DeRisi, J. L., Caccamo, M., Li, Y., Jaffe, D. B., Green, R. E., Haussler, D., Korf, I., and Paten, B. (2011). Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome Res.*, 21(12):2224–2241.

[29] Ebler, J., Ebert, P., Clarke, W. E., Rausch, T., Audano, P. A., Houwaart, T., Mao, Y., Korbel, J. O., Eichler, E. E., Zody, M. C., Dilthey, A. T., and Marschall, T. (2022). Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nat. Genet.*, 54(4):518–525.

[30] Eggertsson, H. P., Jonsson, H., Kristmundsdottir, S., Hjartarson, E., Kehr, B., Masson, G., Zink, F., Hjorleifsson, K. E., Jonasdottir, A., Jonasdottir, A., Jonsdottir, I., Gudbjartsson, D. F., Melsted, P., Stefansson, K., and Halldorsson, B. V. (2017). Graphtyper enables population-scale genotyping using pangenome graphs. *Nat. Genet.*, 49(11):1654–1660.

[31] Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., Dewinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., and Turner, S. (2009). Real-time DNA sequencing from single polymerase molecules. *Science*, 323(5910):133–138.

[32] Eizenga, J. M., Novak, A. M., Kobayashi, E., Villani, F., Cisar, C., Heumos, S., Hickey, G., Colonna, V., Paten, B., and Garrison, E. (2020). Efficient dynamic variation graphs. *Bioinformatics*.

[33] Falconer, E., Hills, M., Naumann, U., Poon, S. S. S., Chavez, E. A., Sanders, A. D., Zhao, Y., Hirst, M., and Lansdorp, P. M. (2012). DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. *Nat. Methods*, 9(11):1107–1112.

[34] Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Min Jou, W., Molemans, F., Raeymaekers, A., Van den Berghe, A., Volckaert, G., and Ysebaert, M. (1976). Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature*, 260(5551):500–507.

[35] Formenti, G., Abueg, L., Brajuka, A., Brajuka, N., Gallardo-Alba, C., Giani, A., Fedrigo, O., and Jarvis, E. D. (2022). Gfastats: conversion, evaluation and manipulation of genome sequences using assembly graphs. *Bioinformatics*, 38(17):4214–4216.

[36] Fromont-Racine, M., Rain, J. C., and Legrain, P. (1997). Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. *Nat. Genet.*, 16(3):277–282.

[37] Fusari, C. M., Kooke, R., Lauxmann, M. A., Annunziata, M. G., Enke, B., Hoehne, M., Krohn, N., Becker, F. F. M., Schlereth, A., Sulpice, R., Stitt, M., and Keurentjes, J. J. B. (2017). Genome-wide association mapping reveals that specific and pleiotropic regulatory mechanisms fine-tune central metabolism and growth in arabidopsis. *Plant Cell*, 29(10):2349–2373.

[38] Garg, S., Rautiainen, M., Novak, A. M., Garrison, E., Durbin, R., and Marschall, T. (2018). A graph-based approach to diploid genome assembly. *Bioinformatics*, 34(13):i105–i114.

[39] Garrison, E., Guarracino, A., Heumos, S., Villani, F., Bao, Z., Tattini, L., Hagmann, J., Vorbrugg, S., Marco-Sola, S., Kubica, C., Ashbrook, D. G., Thorell, K., Rusholme-Pilcher, R. L., Liti, G., Rudbeck, E., Nahnsen, S., Yang, Z., Moses, M. N., Nobrega, F. L., Wu, Y., Chen, H., de Ligt, J., Sudmant, P. H., Soranzo, N., Colonna, V., Williams, R. W., and Prins, P. (2023). Building pangenome graphs. *bioRxiv*, page 2023.04.05.535718.

[40] Garrison, E. and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv [q-bio.GN]*.

[41] Garrison, E., Sirén, J., Novak, A. M., Hickey, G., Eizenga, J. M., Dawson, E. T., Jones, W., Garg, S., Markello, C., Lin, M. F., Paten, B., and Durbin, R. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.*, 36(9):875–879.

[42] Gaurav, K., Arora, S., Silva, P., Sánchez-Martín, J., Horsnell, R., Gao, L., Brar, G. S., Widrig, V., John Raupp, W., Singh, N., Wu, S., Kale, S. M., Chinoy, C., Nicholson, P., Quiroz-Chávez, J., Simmonds, J., Hayta, S., Smedley, M. A., Harwood, W., Pearce, S., Gilbert, D., Kangara, N., Gardener, C., Forner-Martínez, M., Liu, J., Yu, G., Boden, S. A., Pascucci, A., Ghosh, S., Hafeez, A. N., O'Hara, T., Waites, J., Cheema, J., Steuernagel, B., Patpour, M., Justesen, A. F., Liu, S., Rudd, J. C., Avni, R., Sharon, A., Steiner, B., Kirana, R. P., Buerstmayr, H., Mehrabi, A. A., Nasyrova, F. Y., Chayut, N., Matny, O., Steffenson, B. J., Sandhu, N., Chhuneja, P., Lagudah, E., Elkot, A. F., Tyrrell, S., Bian, X., Davey, R. P., Simonsen, M., Schauser, L., Tiwari, V. K., Randy Kutcher, H., Hucl, P., Li, A., Liu, D.-C., Mao, L., Xu, S., Brown-Guedira, G., Faris, J., Dvorak, J., Luo, M.-C., Krasileva, K., Lux, T., Artmeier, S., Mayer, K. F. X., Uauy, C., Mascher, M., Bentley, A. R., Keller, B., Poland, J., and Wulff, B. B. H. (2022). Population genomic analysis of aegilops tauschii identifies targets for bread wheat improvement. *Nat. Biotechnol.*, 40(3):422–431.

[43] Goel, M., Sun, H., Jiao, W.-B., and Schneeberger, K. (2019). SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.*, 20(1):277.

[44] Goff, S. A., Ricke, D., Lan, T.-H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., Hadley, D., Hutchison, D., Martin, C., Katagiri, F., Lange, B. M., Moughamer, T., Xia, Y., Budworth, P., Zhong, J., Miguel, T., Paszkowski, U., Zhang, S., Colbert, M., Sun, W.-L., Chen, L., Cooper, B., Park, S., Wood, T. C., Mao, L., Quail, P., Wing, R., Dean, R., Yu, Y., Zharkikh, A., Shen, R., Sahasrabudhe, S., Thomas, A., Cannings, R., Gutin, A., Pruss, D., Reid, J., Tavtigian, S., Mitchell, J., Eldredge, G., Scholl, T., Miller, R. M., Bhatnagar, S., Adey, N., Rubano, T., Tusneem, N., Robinson, R., Feldhaus, J., Macalma, T., Oliphant, A., and Briggs, S. (2002). A draft sequence of the rice genome (oryza sativa L. ssp. japonica). *Science*, 296(5565):92–100.

[45] Golicz, A. A., Batley, J., and Edwards, D. (2016). Towards plant pangenomics. *Plant Biotechnol. J.*, 14(4):1099–1105.

[46] Guarracino, A., Heumos, S., Nahnsen, S., Prins, P., and Garrison, E. (2022). ODGI: understanding pangenome graphs. *Bioinformatics*, 38(13):3319–3326.

[47] Hansson, B., Kawabe, A., Preuss, S., Kuittinen, H., and Charlesworth, D. (2006). Comparative gene mapping in arabidopsis lyrata chromosomes 1 and 2 and the corresponding

a. thaliana chromosome 1: recombination rates, rearrangements and centromere location. *Genet. Res.*, 87(2):75–85.

[48] Harris, T. D., Buzby, P. R., Babcock, H., Beer, E., Bowers, J., Braslavsky, I., Causey, M., Colonell, J., Dimeo, J., Efcavitch, J. W., Giladi, E., Gill, J., Healy, J., Jarosz, M., Lapen, D., Moulton, K., Quake, S. R., Steinmann, K., Thayer, E., Tyurina, A., Ward, R., Weiss, H., and Xie, Z. (2008). Single-molecule DNA sequencing of a viral genome. *Science*, 320(5872):106–109.

[49] He, C., Washburn, J. D., Hao, Y., Zhang, Z., Yang, J., and Liu, S. (2021). Trait association and prediction through integrative K-mer analysis. *bioRxiv*, page 2021.11.17.468725.

[50] He, Q., Tang, S., Zhi, H., Chen, J., Zhang, J., Liang, H., Alam, O., Li, H., Zhang, H., Xing, L., Li, X., Zhang, W., Wang, H., Shi, J., Du, H., Wu, H., Wang, L., Yang, P., Xing, L., Yan, H., Song, Z., Liu, J., Wang, H., Tian, X., Qiao, Z., Feng, G., Guo, R., Zhu, W., Ren, Y., Hao, H., Li, M., Zhang, A., Guo, E., Yan, F., Li, Q., Liu, Y., Tian, B., Zhao, X., Jia, R., Feng, B., Zhang, J., Wei, J., Lai, J., Jia, G., Purugganan, M., and Diao, X. (2023). A graph-based genome and pan-genome variation of the model plant setaria. *Nat. Genet.*, 55(7):1232–1242.

[51] Heumos, S., Guarracino, A., Schmelzle, J.-N. M., Li, J., Zhang, Z., Hagmann, J., Nahnsen, S., Prins, P., and Garrison, E. (2023). Pangenome graph layout by path-guided stochastic gradient descent. *bioRxiv*.

[52] Hickey, G., Heller, D., Monlong, J., Sibbesen, J. A., Sirén, J., Eizenga, J., Dawson, E. T., Garrison, E., Novak, A. M., and Paten, B. (2020). Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biol.*, 21(1):35.

[53] Hickey, G., Monlong, J., Ebler, J., Novak, A. M., Eizenga, J. M., Gao, Y., Human Pangenome Reference Consortium, Marschall, T., Li, H., and Paten, B. (2023). Pangenome graph construction from genome alignments with minigraph-cactus. *Nat. Biotechnol.*

[54] Hirsch, C. N., Foerster, J. M., Johnson, J. M., Sekhon, R. S., Muttoni, G., Vaillancourt, B., Peñagaricano, F., Lindquist, E., Pedraza, M. A., Barry, K., de Leon, N., Kaeppler, S. M., and Buell, C. R. (2014). Insights into the maize pan-genome and pan-transcriptome. *Plant Cell*, 26(1):121–135.

[55] Hosmani, P. S., Flores-Gonzalez, M., van de Geest, H., Maumus, F., Bakker, L. V., Schijlen, E., van Haarst, J., Cordewener, J., Sanchez-Perez, G., Peters, S., Fei, Z., Giovannoni, J. J., Mueller, L. A., and Saha, S. (2019). An improved de novo assembly and annotation of the tomato reference genome using single-molecule sequencing, hi-C proximity ligation and optical maps. *bioRxiv*, page 767764.

[56] Hu, T. T., Pattyn, P., Bakker, E. G., Cao, J., Cheng, J.-F., Clark, R. M., Fahlgren, N., Fawcett, J. A., Grimwood, J., Gundlach, H., Haberer, G., Hollister, J. D., Ossowski, S., Ottilar, R. P., Salamov, A. A., Schneeberger, K., Spannagl, M., Wang, X., Yang, L., Nasrallah, M. E., Bergelson, J., Carrington, J. C., Gaut, B. S., Schmutz, J., Mayer, K. F. X., Van de Peer, Y., Grigoriev, I. V., Nordborg, M., Weigel, D., and Guo, Y.-L. (2011). The arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat. Genet.*, 43(5):476–481.

[57] Igolkina, A. A., Bezlepsky, A. D., and Nordborg, M. (2025). Pannagram: unbiased pangenome alignment and the mobilome calling. *bioRxiv*, page 2025.02.07.637071.

[58] Igolkina, A. A., Vorbrugg, S., Rabanal, F. A., Liu, H.-J., Ashkenazy, H., Kornienko, A. E., Fitz, J., Collenberg, M., Kubica, C., Morales, A. M., Jaegle, B., Wrightsman, T., Voloshin, V., Llaca, V., Nizhynska, V., Reichardt, I., Lanz, C., Bemm, F., Flood, P. J., Nemomissa, S., Hancock, A., Guo, Y.-L., Kersey, P., Weigel, D., and Nordborg, M. (2024). Towards an unbiased characterization of genetic polymorphism. *Genomics*, (biorxiv;2024.05.30.596703v1):2024.05.30.596703.

[59] Iqbal, Z., Caccamo, M., Turner, I., Flicek, P., and McVean, G. (2012). De novo assembly and genotyping of variants using colored de bruijn graphs.

[60] Jaillard, M., Lima, L., Tournoud, M., Mahé, P., van Belkum, A., Lacroix, V., and Jacob, L. (2018). A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events. *PLoS Genet.*, 14(11):e1007758.

[61] Jaillard, M., Tournoud, M., Lima, L., Lacroix, V., Veyrieras, J.-B., and Jacob, L. (2017). Representing genetic determinants in bacterial GWAS with compacted de bruijn graphs. *bioRxiv*, page 113563.

[62] Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., Tyson, J. R., Beggs, A. D., Dilthey, A. T., Fiddes, I. T., Malla, S., Marriott, H., Nieto, T., O'Grady, J., Olsen, H. E., Pedersen, B. S., Rhie, A., Richardson, H., Quinlan, A. R., Snutch, T. P., Tee, L., Paten, B., Phillippy, A. M., Simpson, J. T., Loman, N. J., and Loose, M. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.*, 36(4):338–345.

[63] Jayakodi, M., Lu, Q., Pidon, H., Rabanus-Wallace, M. T., Bayer, M., Lux, T., Guo, Y., Jaegle, B., Badea, A., Bekele, W., Brar, G. S., Braune, K., Bunk, B., Chalmers, K. J., Chapman, B., Jørgensen, M. E., Feng, J.-W., Feser, M., Fiebig, A., Gundlach, H., Guo, W., Haberer, G., Hansson, M., Himmelbach, A., Hoffie, I., Hoffie, R. E., Hu, H., Isobe, S., König, P., Kale, S. M., Kamal, N., Keeble-Gagnère, G., Keller, B., Knauft, M., Koppolu, R., Krattinger, S. G., Kumlehn, J., Langridge, P., Li, C., Marone, M. P., Maurer, A., Mayer, K. F. X., Melzer, M., Muehlbauer, G. J., Murozuka, E., Padmarasu, S., Perovic, D., Pillen, K., Pin, P. A., Pozniak, C. J., Ramsay, L., Pedas, P. R., Rutten, T., Sakuma, S., Sato, K., Schüler, D., Schmutzer, T., Scholz, U., Schreiber, M., Shirasawa, K., Simpson, C., Skadhauge, B., Spannagl, M., Steffenson, B. J., Thomsen, H. C., Tibbits, J. F., Nielsen, M. T. S., Trautewig, C., Vequaud, D., Voss, C., Wang, P., Waugh, R., Westcott, S., Rasmussen, M. W., Zhang, R., Zhang, X.-Q., Wicker, T., Dockter, C., Mascher, M., and Stein, N. (2024). Structural variation in the pangenome of wild and domesticated barley. *Nature*, 636(8043):654–662.

[64] Jiao, W.-B. and Schneeberger, K. (2017). The impact of third generation genomic technologies on plant genome assembly. *Curr. Opin. Plant Biol.*, 36:64–70.

[65] Jiao, W.-B. and Schneeberger, K. (2020). Chromosome-level assemblies of multiple arabidopsis genomes reveal hotspots of rearrangements with altered evolutionary dynamics. *Nat. Commun.*, 11(1):989.

[66] Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., and Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723.

[67] Kasianowicz, J. J., Brandin, E., Branton, D., and Deamer, D. W. (1996). Characterization of individual polynucleotide molecules using a membrane channel. *Proc. Natl. Acad. Sci. U. S. A.*, 93(24):13770–13773.

[68] Kent, W. J. and Haussler, D. (2001). Assembly of the working draft of the human genome with GigAssembler. *Genome Res.*, 11(9):1541–1548.

[69] Koren, S., Rhie, A., Walenz, B. P., Dilthey, A. T., Bickhart, D. M., Kingan, S. B., Hiendleder, S., Williams, J. L., Smith, T. P. L., and Phillippy, A. M. (2018). De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.*

[70] Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.*, 27(5):722–736.

[71] Korobeynikov, A. (2025). BandageNG: a bioinformatics application for navigating de novo assembly graphs easily.

[72] Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, Y., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L.,

Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., Szustakowki, J., and International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

[73] Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nat. Methods*, 9(4):357–359.

[74] Leonard, A. S., Crysnanto, D., Mapel, X. M., Bhati, M., and Pausch, H. (2023). Graph construction method impacts variation representation and analyses in a bovine super-pangenome. *Genome Biol.*, 24(1):124.

[75] Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100.

[76] Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760.

[77] Li, H., Feng, X., and Chu, C. (2020). The design and construction of reference pangenome graphs with minigraph. *Genome Biol.*, 21(1):265.

[78] Li, H. and Rong, J. (2020). Bedtk: Finding interval overlap with implicit interval tree. page 2020.07.07.190744.

[79] Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., Li, S., Yang, H., Wang, J., and Wang, J. (2010). De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.*, 20(2):265–272.

[80] Li, Y. and Lin, Y. (2020). Kmer2SNP: reference-free SNP calling from raw reads based on matching. *Cold Spring Harbor Laboratory*, page 2020.05.17.100305.

[81] Liao, W.-W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J. K., Monlong, J., Abel, H. J., Buonaiuto, S., Chang, X. H., Cheng, H., Chu, J., Colonna, V., Eizenga, J. M., Feng, X., Fischer, C., Fulton, R. S., Garg, S., Groza, C., Guarracino, A., Harvey, W. T., Heumos, S., Howe, K., Jain, M., Lu, T.-Y., Markello, C., Martin, F. J., Mitchell, M. W., Munson, K. M., Mwaniki, M. N., Novak, A. M., Olsen, H. E., Pesout, T., Porubsky, D., Prins, P., Sibbesen, J. A., Sirén, J., Tomlinson, C., Villani, F., Vollger, M. R., Antonacci-Fulton, L. L., Baid, G., Baker, C. A., Belyaeva, A., Billis, K., Carroll, A., Chang, P.-C., Cody, S., Cook, D. E., Cook-Deegan, R. M., Cornejo, O. E., Diekhans, M., Ebert, P., Fairley, S., Fedrigo, O., Felsenfeld, A. L., Formenti, G., Frankish, A., Gao, Y., Garrison, N. A., Giron, C. G., Green, R. E., Haggerty, L., Hoekzema, K., Hourlier, T., Ji, H. P., Kenny, E. E., Koenig, B. A., Kolesnikov, A., Korbel, J. O., Kordosky, J., Koren,

S., Lee, H., Lewis, A. P., Magalhães, H., Marco-Sola, S., Marijon, P., McCartney, A., McDaniel, J., Mountcastle, J., Nattestad, M., Nurk, S., Olson, N. D., Popejoy, A. B., Puiu, D., Rautiainen, M., Regier, A. A., Rhie, A., Sacco, S., Sanders, A. D., Schneider, V. A., Schultz, B. I., Shafin, K., Smith, M. W., Sofia, H. J., Abou Tayoun, A. N., Thibaud-Nissen, F., Tricomi, F. F., Wagner, J., Walenz, B., Wood, J. M. D., Zimin, A. V., Bourque, G., Chaisson, M. J. P., Flicek, P., Phillippy, A. M., Zook, J. M., Eichler, E. E., Haussler, D., Wang, T., Jarvis, E. D., Miga, K. H., Garrison, E., Marschall, T., Hall, I. M., Li, H., and Paten, B. (2023). A draft human pangenome reference. *Nature*, 617(7960):312–324.

[82] Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., and Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293.

[83] Lightfoot, D. J., Jarvis, D. E., Ramaraj, T., Lee, R., Jellen, E. N., and Maughan, P. J. (2017). Single-molecule sequencing and hi-C-based proximity-guided assembly of amaranth (amaranthus hypochondriacus) chromosomes provide insights into genome evolution. *BMC Biol.*, 15(1):74.

[84] Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., Zhou, G.-A., Zhang, H., Liu, Z., Shi, M., Huang, X., Li, Y., Zhang, M., Wang, Z., Zhu, B., Han, B., Liang, C., and Tian, Z. (2020). Pan-genome of wild and cultivated soybeans. *Cell*, 182(1):162–176.e13.

[85] Luo, X., Kang, X., and Schönhuth, A. (2021). phasebook: haplotype-aware de novo assembly of diploid genomes from long reads. *Genome Biol.*, 22(1):299.

[86] Lysak, M. A., Berr, A., Pecinka, A., Schmidt, R., McBreen, K., and Schubert, I. (2006). Mechanisms of chromosome number reduction in arabidopsis thaliana and related brassicaceae species. *Proc. Natl. Acad. Sci. U. S. A.*, 103(13):5224–5229.

[87] Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, 9(1):387–402.

[88] Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L. I., Jarvie, T. P., Jirage, K. B., Kim, J.-B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F., and Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380.

[89] Martin, G., Baurens, F.-C., Droc, G., Rouard, M., Cenci, A., Kilian, A., Hastie, A., Doležel, J., Aury, J.-M., Alberti, A., Carreel, F., and D'Hont, A. (2016). Improvement of the banana "musa acuminata" reference sequence using NGS data and semi-automated bioinformatics methods. *BMC Genomics*, 17(1):243.

[90] Mascher, M., Wicker, T., Jenkins, J., Plott, C., Lux, T., Koh, C. S., Ens, J., Gundlach, H., Boston, L. B., Tulpová, Z., Holden, S., Hernández-Pinzón, I., Scholz, U., Mayer, K. F. X., Spannagl, M., Pozniak, C. J., Sharpe, A. G., Šimková, H., Moscou, M. J., Grimwood, J., Schmutz, J., and Stein, N. (2021). Long-read sequence assembly: a technical evaluation in barley. *Plant Cell*, 33(6):1888–1906.

[91] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. (2010). The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, 20(9):1297–1303.

[92] Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nat. Rev. Genet.*, 11(1):31–46.

[93] Miescher, F. (1871). *Ueber die chemische Zusammensetzung der Eiterzellen*, volume 4.

[94] Miga, K. H., Koren, S., Rhie, A., Vollger, M. R., Gershman, A., Bzikadze, A., Brooks, S., Howe, E., Porubsky, D., Logsdon, G. A., Schneider, V. A., Potapova, T., Wood, J., Chow, W., Armstrong, J., Fredrickson, J., Pak, E., Tigyi, K., Kremitzki, M., Markovic, C., Maduro, V., Dutra, A., Bouffard, G. G., Chang, A. M., Hansen, N. F., Wilfert, A. B., Thibaud-Nissen, F., Schmitt, A. D., Belton, J.-M., Selvaraj, S., Dennis, M. Y., Soto, D. C., Sahasrabudhe, R., Kaya, G., Quick, J., Loman, N. J., Holmes, N., Loose, M., Surti, U., Risques, R. A., Graves Lindsay, T. A., Fulton, R., Hall, I., Paten, B., Howe, K., Timp, W., Young, A., Mullikin, J. C., Pevzner, P. A., Gerton, J. L., Sullivan, B. A., Eichler, E. E., and Phillippy, A. M. (2020). Telomere-to-telomere assembly of a complete human X chromosome. *Nature*, 585(7823):79–84.

[95] Mikheyev, A. S. and Tin, M. M. Y. (2014). A first look at the oxford nanopore MinION sequencer. *Mol. Ecol. Resour.*, 14(6):1097–1102.

[96] Milia, S., Leonard, A. S., Mapel, X. M., Bernal Ulloa, S. M., Drögemüller, C., and Pausch, H. (2025). Taurine pangenome uncovers a segmental duplication upstream of KIT associated with depigmentation in white-headed cattle. *Genome Res.*, 35(4):1041–1052.

[97] Mitra, R. D. and Church, G. M. (1999). In situ localized amplification and contact replication of many individual DNA molecules. *Nucleic Acids Res.*, 27(24):e34.

[98] Myers, E. W. (2005). The fragment assembly string graph. *Bioinformatics*, 21 Suppl 2:ii79–85.

[99] Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., Kravitz, S. A., Mobarry, C. M., Reinert, K. H., Remington, K. A., Anson, E. L., Bolanos, R. A., Chou, H. H., Jordan, C. M., Halpern, A. L., Lonardi, S., Beasley, E. M., Brandon, R. C., Chen, L., Dunn, P. J., Lai, Z., Liang, Y., Nusskern, D. R., Zhan, M., Zhang, Q., Zheng, X., Rubin, G. M., Adams, M. D., and Venter, J. C. (2000). A whole-genome assembly of drosophila. *Science*, 287(5461):2196–2204.

[100] 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., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., and Köster, J. (2021). Sustainable data analysis with snakemake. *F1000Res.*, 10:33.

[101] Naish, M., Alonge, M., Wlodzimierz, P., Tock, A. J., Abramson, B. W., Schmücker, A., Mandáková, T., Jamge, B., Lambing, C., Kuo, P., Yelina, N., Hartwick, N., Colt, K., Smith, L. M., Ton, J., Kakutani, T., Martienssen, R. A., Schneeberger, K., Lysak, M. A., Berger, F., Bousios, A., Michael, T. P., Schatz, M. C., and Henderson, I. R. (2021). The genetic and epigenetic landscape of the arabidopsis centromeres. *Science*, 374(6569):eabi7489.

[102] Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., Caldas, G. V., Chen, N.-C., Cheng, H., Chin, C.-S., Chow, W., de Lima, L. G., Dishuck, P. C., Durbin, R., Dvorkina, T., Fiddes, I. T., Formenti, G., Fulton, R. S., Fungtammasan, A., Garrison, E., Grady, P. G. S., Graves-Lindsay, T. A., Hall, I. M., Hansen, N. F., Hartley, G. A., Haukness, M., Howe, K., Hunkapiller, M. W., Jain, C., Jain, M., Jarvis, E. D., Kerpedjiev, P., Kirsche, M., Kolmogorov, M., Korlach, J., Kremitzki, M., Li, H., Maduro, V. V., Marschall, T., McCartney, A. M., McDaniel, J., Miller, D. E., Mullikin, J. C., Myers, E. W., Olson, N. D., Paten, B., Peluso, P., Pevzner, P. A., Porubsky, D., Potapova, T., Rogaev, E. I., Rosenfeld, J. A., Salzberg, S. L., Schneider, V. A., Sedlazeck, F. J., Shafin, K., Shew, C. J., Shumate, A., Sims, Y., Smit, A. F. A., Soto, D. C., Sović, I., Storer, J. M., Streets, A., Sullivan, B. A., Thibaud-Nissen, F., Torrance, J., Wagner, J., Walenz, B. P., Wenger, A., Wood, J. M. D., Xiao, C., Yan, S. M., Young, A. C., Zarate, S., Surti, U., McCoy, R. C., Dennis, M. Y., Alexandrov, I. A., Gerton, J. L., O'Neill, R. J., Timp, W., Zook, J. M., Schatz, M. C., Eichler, E. E., Miga, K. H., and Phillippy, A. M. (2022). The complete sequence of a human genome. *Science*, 376(6588):44–53.

[103] O'Donnell, S., Yue, J.-X., Saada, O. A., Agier, N., Caradec, C., Cokelaer, T., De Chiara, M., Delmas, S., Dutreux, F., Fournier, T., Friedrich, A., Kornobis, E., Li, J., Miao, Z., Tattini, L., Schacherer, J., Liti, G., and Fischer, G. (2023). Telomere-to-telomere assemblies of 142 strains characterize the genome structural landscape in saccharomyces cerevisiae. *Nat. Genet.*, 55(8):1390–1399.

[104] Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., and Phillippy, A. M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.*, 17(1):132.

[105] Onodera, T., Sadakane, K., Shibuya, T., Darling, A., and Stoye, J. (2013). Algorithms in bioinformatics.

[106] Osoegawa, K., Tateno, M., Woon, P. Y., Frengen, E., Mammoser, A. G., Catanese, J. J., Hayashizaki, Y., and de Jong, P. J. (2000). Bacterial artificial chromosome libraries for mouse sequencing and functional analysis. *Genome Res.*, 10(1):116–128.

[107] Ossowski, S., Schneeberger, K., Clark, R. M., Lanz, C., Warthmann, N., and Weigel, D. (2008). Sequencing of natural strains of arabidopsis thaliana with short reads. *Genome Res.*, 18(12):2024–2033.

[108] Pangenome Reference, H. and others (2023). Pangenome graph construction from genome alignments with minigraph-cactus. *Nature*.

[109] Paten, B., Eizenga, J. M., Rosen, Y. M., Novak, A. M., Garrison, E., and Hickey, G. (2018). Superbubbles, ultrabubbles, and cacti. *J. Comput. Biol.*, 25(7):649–663.

[110] Paten, B., Novak, A. M., Eizenga, J. M., and Garrison, E. (2017). Genome graphs and the evolution of genome inference. *Genome Res.*, 27(5):665–676.

[111] Peter, J., De Chiara, M., Friedrich, A., Yue, J.-X., Pflieger, D., Bergström, A., Sigwalt, A., Barre, B., Freel, K., Llored, A., Cruaud, C., Labadie, K., Aury, J.-M., Istace, B., Lebrigand, K., Barbry, P., Engelen, S., Lemainque, A., Wincker, P., Liti, G., and Schacherer, J. (2018). Genome evolution across 1,011 saccharomyces cerevisiae isolates. *Nature*, 556(7701):339–344.

[112] Pevzner, P. A., Tang, H., and Waterman, M. S. (2001). An eulerian path approach to DNA fragment assembly. *Proc. Natl. Acad. Sci. U. S. A.*, 98(17):9748–9753.

[113] Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P. T., Gross, S. S., Dorfman, L., McLean, C. Y., and DePristo, M. A. (2018). A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.*, 36(10):983–987.

[114] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., and Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, 81(3):559–575.

[115] Quick, J., Loman, N. J., Duraffour, S., Simpson, J. T., Severi, E., Cowley, L., Bore, J. A., Koundouno, R., Dudas, G., Mikhail, A., Ouédraogo, N., Afrough, B., Bah, A., Baum, J. H., Becker-Ziaja, B., Boettcher, J.-P., Cabeza-Cabrerizo, M., Camino-Sanchez, A., Carter, L. L., Doerrbecker, J., Enkirch, T., Dorival, I. G. G., Hetzelt, N., Hinzmann, J., Holm, T., Kafetzopoulou, L. E., Koropogui, M., Kosgey, A., Kuisma, E., Logue, C. H., Mazzarelli, A., Meisel, S., Mertens, M., Michel, J., Ngabo, D., Nitzsche, K., Pallash, E., Patrono, L. V., Portmann, J., Repits, J. G., Rickett, N. Y., Sachse, A., Singethan, K., Vitoriano, I., Yemanaberhan, R. L., Zekeng, E. G., Trina, R., Bello, A., Sall, A. A., Faye, O., Faye, O., Magassouba, N., Williams, C. V., Amburgey, V., Winona, L., Davis, E., Gerlach, J., Washington, F., Monteil, V., Jourdain, M., Bererd, M., Camara, A., Somlare, H., Camara, A., Gerard, M., Bado, G., Baillet, B., Delaune, D., Nebie, K. Y., Diarra, A., Savane, Y., Pallawo, R. B., Gutierrez, G. J., Milhano, N., Roger, I., Williams, C. J., Yattara, F., Lewandowski, K., Taylor, J., Rachwal, P., Turner, D., Pollakis, G., Hiscox, J. A., Matthews, D. A., O'Shea, M. K., Johnston, A. M., Wilson, D., Hutley, E., Smit, E., Di Caro, A., Woelfel, R., Stoecker, K., Fleischmann, E., Gabriel, M., Weller, S. A., Koivogui, L., Diallo, B., Keita, S., Rambaut, A., Formenty, P., Gunther, S., and Carroll, M. W. (2016). Real-time, portable genome sequencing for ebola surveillance. *Nature*, 530(7589):228–232.

[116] Rautiainen, M. and Marschall, T. (2020). GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome Biol.*, 21(1):253.

[117] Roberts, R. J., Carneiro, M. O., and Schatz, M. C. (2013). The advantages of SMRT sequencing. *Genome Biol.*, 14(7):1–4.

[118] Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., Leamon, J. H., Johnson, K., Milgrew, M. J., Edwards, M., Hoon, J., Simons, J. F., Marran, D., Myers, J. W., Davidson, J. F., Branting, A., Nobile, J. R., Puc, B. P., Light, D., Clark, T. A., Huber,

M., Branciforte, J. T., Stoner, I. B., Cawley, S. E., Lyons, M., Fu, Y., Homer, N., Sedova, M., Miao, X., Reed, B., Sabina, J., Feierstein, E., Schorn, M., Alanjary, M., Dimalanta, E., Dressman, D., Kasinskas, R., Sokolsky, T., Fidanza, J. A., Namsaraev, E., McKernan, K. J., Williams, A., Roth, G. T., and Bustillo, J. (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475(7356):348–352.

[119] Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, C. A., Hutchison, C. A., Slocombe, P. M., and Smith, M. (1977a). Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*, 265(5596):687–695.

[120] Sanger, F., Nicklen, S., and Coulson, A. R. (1977b). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.*, 74(12):5463–5467.

[121] Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T. A., Minx, P., Reily, A. D., Courtney, L., Kruchowski, S. S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., Rock, S. M., Belter, E., Du, F., Kim, K., Abbott, R. M., Cotton, M., Levy, A., Marchetto, P., Ochoa, K., Jackson, S. M., Gillam, B., Chen, W., Yan, L., Higginbotham, J., Cardenas, M., Waligorski, J., Applebaum, E., Phelps, L., Falcone, J., Kanchi, K., Thane, T., Scimone, A., Thane, N., Henke, J., Wang, T., Ruppert, J., Shah, N., Rotter, K., Hodges, J., Ingenthron, E., Cordes, M., Kohlberg, S., Sgro, J., Delgado, B., Mead, K., Chinwalla, A., Leonard, S., Crouse, K., Collura, K., Kudrna, D., Currie, J., He, R., Angelova, A., Rajasekar, S., Mueller, T., Lomeli, R., Scara, G., Ko, A., Delaney, K., Wissotski, M., Lopez, G., Campos, D., Braidotti, M., Ashley, E., Golser, W., Kim, H., Lee, S., Lin, J., Dujmic, Z., Kim, W., Talag, J., Zuccolo, A., Fan, C., Sebastian, A., Kramer, M., Spiegel, L., Nascimento, L., Zutavern, T., Miller, B., Ambroise, C., Muller, S., Spooner, W., Narechania, A., Ren, L., Wei, S., Kumari, S., Faga, B., Levy, M. J., McMahan, L., Van Buren, P., Vaughn, M. W., Ying, K., Yeh, C.-T., Emrich, S. J., Jia, Y., Kalyanaraman, A., Hsia, A.-P., Barbazuk, W. B., Baucom, R. S., Brutnell, T. P., Carpita, N. C., Chaparro, C., Chia, J.-M., Deragon, J.-M., Estill, J. C., Fu, Y., Jeddeloh, J. A., Han, Y., Lee, H., Li, P., Lisch, D. R., Liu, S., Liu, Z., Nagel, D. H., McCann, M. C., SanMiguel, P., Myers, A. M., Nettleton, D., Nguyen, J., Penning, B. W., Ponnala, L., Schneider, K. L., Schwartz, D. C., Sharma, A., Soderlund, C., Springer, N. M., Sun, Q., Wang, H., Waterman, M., Westerman, R., Wolfgruber, T. K., Yang, L., Yu, Y., Zhang, L., Zhou, S., Zhu, Q., Bennetzen, J. L., Dawe, R. K., Jiang, J., Jiang, N., Presting, G. G., Wessler, S. R., Aluru, S., Martienssen, R. A., Clifton, S. W., McCombie, W. R., Wing, R. A., and Wilson, R. K. (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science*, 326(5956):1112–1115.

[122] Schneeberger, K., Ossowski, S., Lanz, C., Juul, T., Petersen, A. H., Nielsen, K. L., Jørgensen, J.-E., Weigel, D., and Andersen, S. U. (2009). SHOREmap: simultaneous mapping and mutation identification by deep sequencing. *Nat. Methods*, 6(8):550–551.

[123] Sedlazeck, F. J., Lee, H., Darby, C. A., and Schatz, M. C. (2018). Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat. Rev. Genet.*, 19(6):329–346.

[124] Servin, B. and Stephens, M. (2005). Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet.*, preprint(2007):e114.

[125] Shang, L., Li, X., He, H., Yuan, Q., Song, Y., Wei, Z., Lin, H., Hu, M., Zhao, F., Zhang, C., Li, Y., Gao, H., Wang, T., Liu, X., Zhang, H., Zhang, Y., Cao, S., Yu, X., Zhang, B., Zhang, Y., Tan, Y., Qin, M., Ai, C., Yang, Y., Zhang, B., Hu, Z., Wang, H., Lv, Y., Wang, Y., Ma, J., Wang, Q., Lu, H., Wu, Z., Liu, S., Sun, Z., Zhang, H., Guo, L., Li, Z., Zhou, Y., Li, J., Zhu, Z., Xiong, G., Ruan, J., and Qian, Q. (2022). A super pan-genomic landscape of rice. *Cell Res.*, 32(10):878–896.

[126] Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A., and Waterston, R. H. (2017). DNA sequencing at 40: past, present and future. *Nature*, 550(7676):345–353.

[127] Shendure, J. and Ji, H. (2008). Next-generation DNA sequencing. *Nat. Biotechnol.*, 26(10):1135–1145.

[128] Shizuya, H., Birren, B., Kim, U. J., Mancino, V., Slepak, T., Tachiiri, Y., and Simon, M. (1992). Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in escherichia coli using an F-factor-based vector. *Proc. Natl. Acad. Sci. U. S. A.*, 89(18):8794–8797.

[129] Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M., and Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. *Genome Res.*, 19(6):1117–1123.

[130] Sirén, J., Monlong, J., Chang, X., Novak, A. M., Eizenga, J. M., Markello, C., Sibbesen, J., Hickey, G., Chang, P.-C., Carroll, A., Haussler, D., Garrison, E., and Paten, B. (2020). Genotyping common, large structural variations in 5,202 genomes using pangenomes, the giraffe mapper, and the vg toolkit. *Cold Spring Harbor Laboratory*, page 2020.12.04.412486.

[131] Smith, T. F. and Waterman, M. S. (1981). Comparison of biosequences. *Adv. Appl. Math.*, 2(4):482–489.

[132] Stevenson, K. R., Coolon, J. D., and Wittkopp, P. J. (2013). Sources of bias in measures of allele-specific expression derived from RNA-sequence data aligned to a single reference genome. *BMC Genomics*, 14(1):536.

[133] Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M. H.-Y., Konkel, M. K., Malhotra, A., Stütz, A. M., Shi, X., Casale, F. P., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., Malig, M., Chaisson, M. J. P., Walter, K., Meiers, S., Kashin, S., Garrison, E., Auton, A., Lam, H. Y. K., Mu, X. J., Alkan, C., Antaki, D., Bae, T., Cerveira, E., Chines, P., Chong, Z., Clarke, L., Dal, E., Ding, L., Emery, S., Fan, X., Gujral, M., Kahveci, F., Kidd, J. M., Kong, Y., Lameijer, E.-W., McCarthy, S., Flicek, P., Gibbs, R. A., Marth, G., Mason, C. E., Menelaou, A., Muzny, D. M., Nelson, B. J., Noor, A., Parrish, N. F., Pendleton, M., Quitadamo, A., Raeder, B., Schadt, E. E., Romanovitch, M., Schlattl, A., Sebra, R., Shabalin, A. A., Untergasser, A., Walker, J. A., Wang, M., Yu, F., Zhang, C., Zhang, J., Zheng-Bradley, X., Zhou, W., Zichner, T., Sebat, J., Batzer, M. A., McCarroll, S. A., 1000 Genomes Project Consortium, Mills, R. E., Gerstein, M. B., Bashir, A., Stegle, O., Devine, S. E., Lee, C., Eichler, E. E., and Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571):75–81.

[134] Sutton, G. G., White, O., Adams, M. D., and Kerlavage, A. R. (1995). TIGR assembler: A new tool for assembling large shotgun sequencing projects. *Genome Science and Technology*, 1(1):9–19.

[135] Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L., Durkin, A. S., Deboy, R. T., Davidsen, T. M., Mora, M., Scarselli, M., Margarit y Ros, I., Peterson, J. D., Hauser, C. R., Sundaram, J. P., Nelson, W. C., Madupu, R., Brinkac, L. M., Dodson, R. J., Rosovitz, M. J., Sullivan, S. A., Daugherty, S. C., Haft, D. H., Selengut, J., Gwinn, M. L., Zhou, L., Zafar, N., Khouri, H., Radune, D., Dimitrov, G., Watkins, K., O'Connor, K. J. B., Smith, S., Utterback, T. R., White, O., Rubens, C. E., Grandi, G., Madoff, L. C., Kasper, D. L., Telford, J. L., Wessels, M. R., Rappuoli, R., and Fraser, C. M. (2005). Genome analysis of multiple pathogenic isolates of streptococcus agalactiae: implications for the microbial "pan-genome". *Proc. Natl. Acad. Sci. U. S. A.*, 102(39):13950–13955.

[136] Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigó, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Deslattes Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan,

C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X. (2001). The sequence of the human genome. *Science*, 291(5507):1304–1351.

[137] Voichek, Y. and Weigel, D. (2020). Identifying genetic variants underlying phenotypic variation in plants without complete genomes. *Nat. Genet.*, 52(5):534–540.

[138] Walkowiak, S., Gao, L., Monat, C., Haberer, G., Kassa, M. T., Brinton, J., Ramirez-Gonzalez, R. H., Kolodziej, M. C., Delorean, E., Thambugala, D., Klymiuk, V., Byrns, B., Gundlach, H., Bandi, V., Siri, J. N., Nilsen, K., Aquino, C., Himmelbach, A., Copetti, D., Ban, T., Venturini, L., Bevan, M., Clavijo, B., Koo, D.-H., Ens, J., Wiebe, K., N'Diaye, A., Fritz, A. K., Gutwin, C., Fiebig, A., Fosker, C., Fu, B. X., Accinelli, G. G., Gardner, K. A., Fradgley, N., Gutierrez-Gonzalez, J., Halstead-Nussloch, G., Hatakeyama, M., Koh, C. S., Deek, J., Costamagna, A. C., Fobert, P., Heavens, D., Kanamori, H., Kawaura, K., Kobayashi, F., Krasileva, K., Kuo, T., McKenzie, N., Murata, K., Nabeka, Y., Paape, T., Padmarasu, S., Percival-Alwyn, L., Kagale, S., Scholz, U., Sese, J., Juliana, P., Singh, R., Shimizu-Inatsugi, R., Swarbreck, D., Cockram, J., Budak, H., Tameshige, T., Tanaka, T., Tsuji, H., Wright, J., Wu, J., Steuernagel, B., Small, I., Cloutier, S., Keeble-Gagnère, G., Muehlbauer, G., Tibbets, J., Nasuda, S., Melonek, J., Hucl, P. J., Sharpe, A. G., Clark, M., Legg, E., Bharti, A., Langridge, P., Hall, A., Uauy, C., Mascher, M., Krattinger, S. G., Handa, H., Shimizu, K. K., Distelfeld, A., Chalmers, K., Keller, B., Mayer, K. F. X., Poland, J., Stein, N., McCartney, C. A., Spannagl, M., Wicker, T., and Pozniak, C. J. (2020). Multiple wheat genomes reveal global variation in modern breeding. *Nature*, 588(7837):277–283.

[139] Watson, J. D. and Crick, F. H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738.

[140] Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P.-C., Hall, R. J., Concepcion, G. T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N. D., Töpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C.-S., Phillippy, A. M., Schatz, M. C., Myers, G., DePristo, M. A., Ruan, J., Marschall, T., Sedlazeck, F. J., Zook, J. M., Li, H., Koren, S., Carroll, A., Rank, D. R., and Hunkapiller, M. W. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.*, 37(10):1155–1162.

[141] Wick, R. R., Judd, L. M., and Holt, K. E. (2019). Performance of neural network basecalling tools for oxford nanopore sequencing. *Genome Biol.*, 20(1):129.

[142] Wick, R. R., Schultz, M. B., Zobel, J., and Holt, K. E. (2015). Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics*, 31(20):3350–3352.

[143] Wlodzimierz, P., Rabanal, F. A., Burns, R., Naish, M., Primetis, E., Scott, A., Mandáková, T., Gorringe, N., Tock, A. J., Holland, D., Fritschi, K., Habring, A., Lanz, C., Patel, C., Schlegel, T., Collenberg, M., Mielke, M., Nordborg, M., Roux, F., Shirsekar, G., Alonso-Blanco, C., Lysak, M. A., Novikova, P. Y., Bousios, A., Weigel, D., and Henderson, I. R. (2023). Cycles of satellite and transposon evolution in arabidopsis centromeres. *Nature*, 618(7965):557–565.

[144] Wu, R. (1972). Nucleotide sequence analysis of DNA. *Nat. New Biol.*, 236(68):198–200.

[145] Yu, J., Hu, S., Wang, J., Wong, G. K.-S., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., Cao, M., Liu, J., Sun, J., Tang, J., Chen, Y., Huang, X., Lin, W., Ye, C., Tong, W., Cong, L., Geng, J., Han, Y., Li, L., Li, W., Hu, G., Huang, X., Li, W., Li, J., Liu, Z., Li, L., Liu, J., Qi, Q., Liu, J., Li, L., Li, T., Wang, X., Lu, H., Wu, T., Zhu, M., Ni, P., Han, H., Dong, W., Ren, X., Feng, X., Cui, P., Li, X., Wang, H., Xu, X., Zhai, W., Xu, Z., Zhang, J., He, S., Zhang, J., Xu, J., Zhang, K., Zheng, X., Dong, J., Zeng, W., Tao, L., Ye, J., Tan, J., Ren, X., Chen, X., He, J., Liu, D., Tian, W., Tian, C., Xia, H., Bao, Q., Li, G., Gao, H., Cao, T., Wang, J., Zhao, W., Li, P., Chen, W., Wang, X., Zhang, Y., Hu, J., Wang, J., Liu, S., Yang, J., Zhang, G., Xiong, Y., Li, Z., Mao, L., Zhou, C., Zhu, Z., Chen, R., Hao, B., Zheng, W., Chen, S., Guo, W., Li, G., Liu, S., Tao, M., Wang, J., Zhu, L., Yuan, L., and Yang, H. (2002). A draft sequence of the rice genome (oryza sativa L. ssp. indica). *Science*, 296(5565):79–92.

[146] Zerbino, D. R. and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome Res.*, 18(5):821–829.

[147] Zheng, G. X. Y., Lau, B. T., Schnall-Levin, M., Jarosz, M., Bell, J. M., Hindson, C. M., Kyriazopoulou-Panagiotopoulou, S., Masquelier, D. A., Merrill, L., Terry, J. M., Mudivarti, P. A., Wyatt, P. W., Bharadwaj, R., Makarewicz, A. J., Li, Y., Belgrader, P., Price, A. D., Lowe, A. J., Marks, P., Vurens, G. M., Hardenbol, P., Montesclaros, L., Luo, M., Greenfield, L., Wong, A., Birch, D. E., Short, S. W., Bjornson, K. P., Patel, P., Hopmans, E. S., Wood, C., Kaur, S., Lockwood, G. K., Stafford, D., Delaney, J. P., Wu, I., Ordonez, H. S., Grimes, S. M., Greer, S., Lee, J. Y., Belhocine, K., Giorda, K. M., Heaton, W. H., McDermott, G. P., Bent, Z. W., Meschi, F., Kondov, N. O., Wilson, R., Bernate, J. A., Gauby, S., Kindwall, A., Bermejo, C., Fehr, A. N., Chan, A., Saxonov, S., Ness, K. D., Hindson, B. J., and Ji, H. P. (2016). Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat. Biotechnol.*, 34(3):303–311.

[148] Zhou, B., Wen, S., Wang, L., Jin, L., Li, H., and Zhang, H. (2017). AntCaller: an accurate variant caller incorporating ancient DNA damage. *Mol. Genet. Genomics*, 292(6):1419–1430.

[149] Zhou, Y., Chebotarov, D., Kudrna, D., Llaca, V., Lee, S., Rajasekar, S., Mohammed, N., Al-Bader, N., Sobel-Sorenson, C., Parakkal, P., Arbelaez, L. J., Franco, N., Alexandrov, N., Hamilton, N. R. S., Leung, H., Mauleon, R., Lorieux, M., Zuccolo, A., McNally, K., Zhang, J., and Wing, R. A. (2020). A platinum standard pan-genome resource that represents the population structure of asian rice. *Sci. Data*, 7(1):113.

[150] Zhou, Y., Zhang, Z., Bao, Z., Li, H., Lyu, Y., Zan, Y., Wu, Y., Cheng, L., Fang, Y., Wu, K., Zhang, J., Lyu, H., Lin, T., Gao, Q., Saha, S., Mueller, L., Fei, Z., Städler, T., Xu, S., Zhang, Z., Speed, D., and Huang, S. (2022). Graph pangenome captures missing heritability and empowers tomato breeding. *Nature*, 606(7914):527–534.

[151] Zou, Y.-P., Hou, X.-H., Wu, Q., Chen, J.-F., Li, Z.-W., Han, T.-S., Niu, X.-M., Yang, L., Xu, Y.-C., Zhang, J., et al. (2017). Adaptation of arabidopsis thaliana to the yangtze river basin. *Genome Biol.*, 18:1–11.

Appendix A

Abbreviations & Glossary

Bubble	A closed subgraph structure with upstream and downstream anchor nodes and intermediate nodes representing sequence variation. A broader form of this is called a snarl (Paten et al., 2018).
Core	The portion of a pangenome common to most or all included accessions.
Edge	A connection in a genome graph that links nodes and determines their order.
GFA	Graphical Fragment Assembly – a human-readable file format for representing genome graphs (GFA, 2022).
GO	Gene Ontology – a framework for the standardized representation of gene and gene product attributes across species, describing their functions, biological processes, and cellular locations.
GWAS	Genome-Wide Association Study – a method used to identify genetic variants associated with traits by scanning markers across the genomes of many individuals.
Node	A basic unit in a genome graph that stores a DNA sequence.
Path	A color-coded route through a genome graph, representing an individual genome's sequence.
Pangenome	The complete set of genomic sequences from multiple individuals representing a population.
PAV	Presence-Absence Variation – a type of structural variation where sequences are either present or missing.
Private	Genomic content found in only one or very few individuals in a pangenome.
Soft	Genomic elements in a pangenome that are neither core nor private.

SNP	Single-Nucleotide Polymorphism – a single base change between two DNA sequences.
SV	Structural Variation – large-scale genome changes such as deletions, duplications, and translocations.
TE	Transposable Element – a mobile DNA sequence capable of copying and inserting itself elsewhere in the genome.
Traversal	A specific path through a genome graph from a start to an end node.

Appendix B

Additional tools and technical background

B.1 Gfa-annotate – Connecting graph statistics with annotation

Variation graphs are data structures designed to represent sequences and are often constructed with the specific goal of capturing diversity across populations or species. However, as with reference-based approaches, meaningful biological interpretation requires knowing which genomic regions are represented by specific nodes in the graph. Therefore, linking graph nodes to functional annotations or reference coordinates is a critical step. This enables researchers, especially those without specialized training, to apply graph-based methods to real biological questions in a practical and interpretable manner.

To this end, we developed a lightweight tool called `gfa-annotate`[1], designed to link flat reference-based annotations to graph-based sequence representations. The tool processes standard annotation files (gff) and overlaps them with graph paths, assigning relevant annotation information to each graph node. The output is a plain-text file, with one node per line and annotation data in the second column. This annotation can include feature types/categories (e.g., mRNA, gene) or gene names and IDs, effectively serving as a dictionary that can be parsed by downstream workflows.

Internally, `gfa-annotate` uses genomic ranges to identify overlaps between graph path intervals and annotated reference features. In addition to basic annotation mapping, the tool can quantify the proportion of each node covered by annotation, a critical metric in large, complex graph regions where individual nodes may span multiple features or functional

[1] https://github.com/MoinSebi/gfa_annotate

elements. This annotation linkage enables more accessible interpretation and utility of graph-based genomic data for researchers without deep specialization in graph bioinformatics. To demonstrate the utility of our tool `gfa-annotate`, and its compatibility with downstream tools such as `gretl`, we conducted a small experiment. The objective was to identify rare gene groups, those that are not present in all accessions but are associated with potentially interesting biological functions.

We used `gfa-annotate` to annotate all graph nodes with manually curated Araport11 annotations. This was feasible because the underlying genome (TAIR10), which Araport11 is based on, was included in the 1001+ genomes graph. For each gene, we extracted all corresponding graph nodes that the gene path traversed (e.g., `AT1G23430` $\rightarrow$ `node1, node2, node3`).

Next, we applied `gretl` to compute node-specific statistics, such as average node length, similarity, depth, and the number of nodes. We intersected these node-level statistics with gene paths, resulting in a gene-wise summary table where each row corresponds to a gene and each column represents a graph-derived statistic.

This summary enabled downstream exploratory analysis, such as identifying genes with unusual graph (population) characteristics. In a further step, we aimed to pinpoint genes that may be under selection pressure, i.e., those found in nearly all accessions or in only a few. This could indicate either essential genes or novel gene candidates (e.g., orthologs or paralogs).

To better understand the functional relevance of these genes, and remove technical artifacts, we summarized genes in their Gene Ontology(GO) groups. Our results showed that several GO terms exhibited 100% sequence similarity across all accessions (Figure B.1). These terms were associated with essential biological functions, whose loss would likely be lethal to the organism. In contrast, terms with low similarity often corresponded to sparsely represented genes, possibly reflecting lineage-specific or conditionally expressed functions (Figure B.2).

While some approaches encode annotations directly as paths within graph files (e.g., in the GFA format), such methods are often inflexible and difficult to use for downstream analysis, especially when dealing with large, diverse annotations or requiring custom filtering and statistics. Our approach decouples annotation from graph structure while preserving a clear and accessible link between the two.

Ultimately, by combining graph-aware statistics, functional annotation, and comparative analyses across accessions, this framework enables the discovery of novel genetic variation, insights into evolutionary dynamics, and a deeper understanding of gene function in popula-

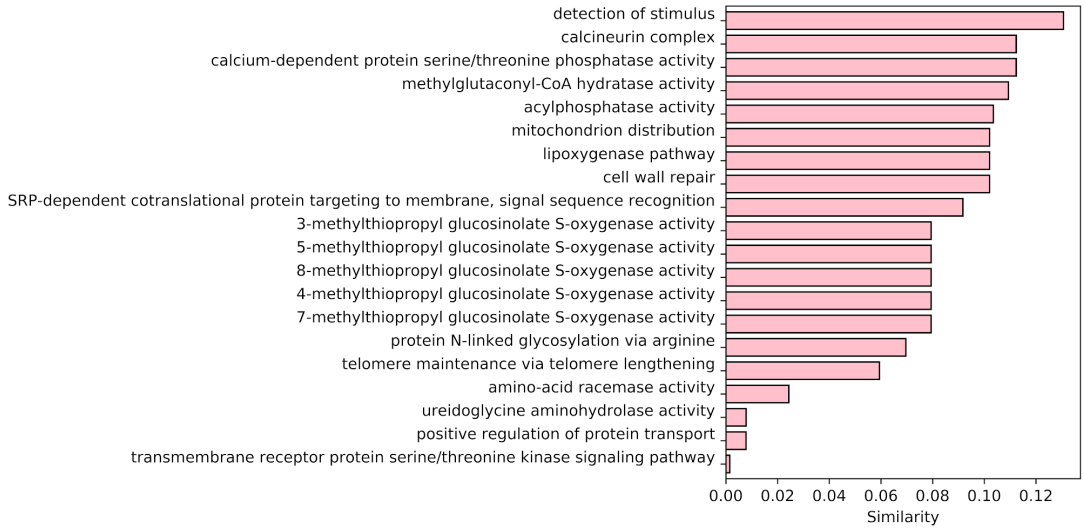


Fig. B.1 **OG with high similarity in *Arabidopsis thaliana***

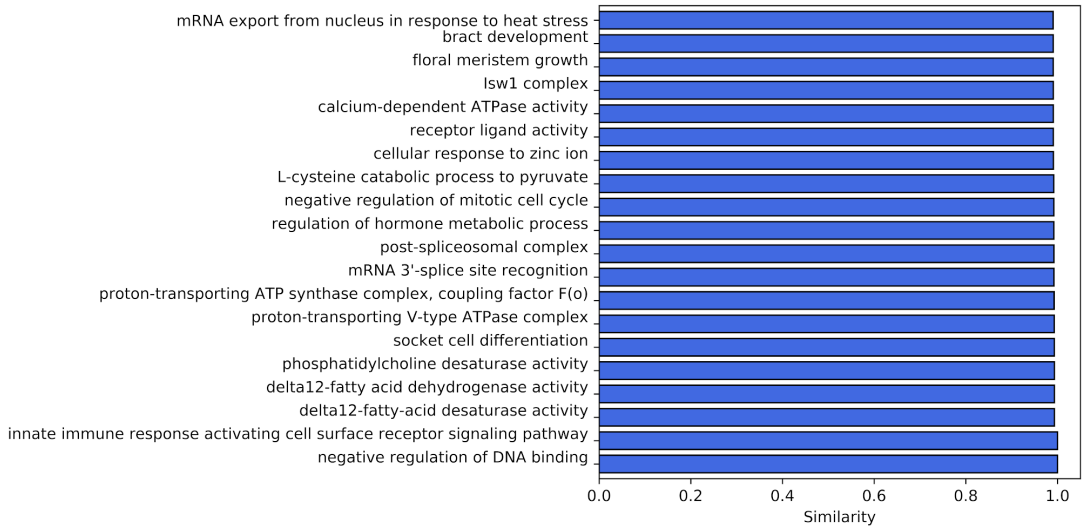


Fig. B.2 **OG with low similarity in *Arabidopsis thaliana***

tion contexts. As genome graphs become more widespread, such tools will be essential for realizing their full potential.

B.2 `Packing` tool

Resequencing remains a important part of population genetics, now focusing on long-reads instead of short-reads, making it essential to develop methods capable of comparing newly

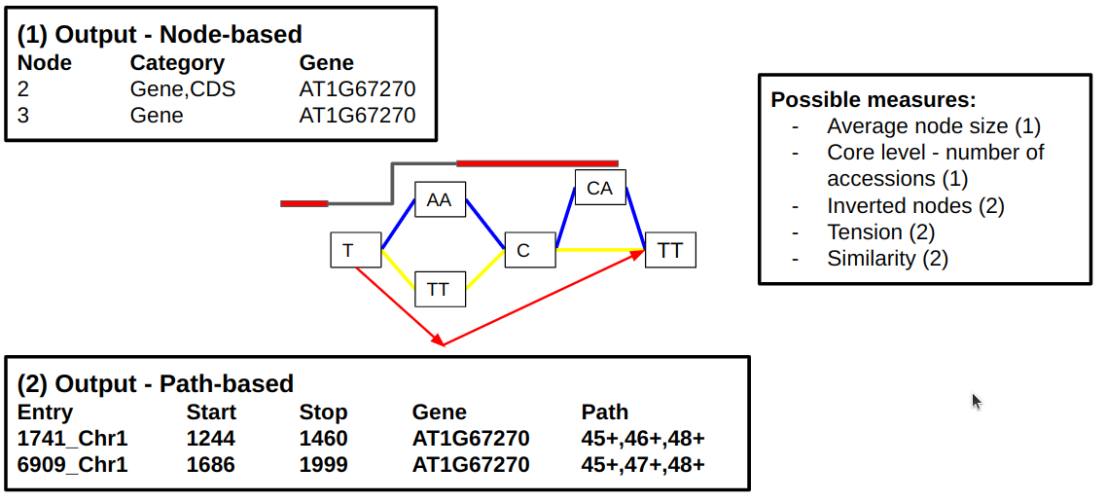


Fig. B.3 **gfa-annotate schematic overview.**

sequenced genomes against existing genome graphs. Whether through direct alignment or alternative strategies, the future direction of genome graph development will depend heavily on how effectively new sequences can be integrated and analyzed.

Currently, alignments to genome graphs are typically stored in the Graph Alignment Format (GAF), a plain-text format similar to PAF. These alignments can be used for a range of downstream analyses, with variant calling being one of the most prominent applications in population genomics. While detecting novel variation remains a major goal, methods for incorporating these variants directly into graphs are not yet standardized.

To address challenges in alignment-based graph analysis, we developed `packing`, a tool that enables efficient storage of coverage profiles derived from GAF alignments. Our format uses compact coverage vectors tailored to a specific genome graph and alignment. By implementing a custom binary structure, we achieve up to 95% reduction in storage size compared to gzip-compressed text formats.

Each coverage vector file is graph-specific and reflects the alignment strategy used. Together with a graph-specific index, this format allows nearly lossless representation of per-node coverage, supporting values up to 65,536. We also provide normalization and thresholding techniques to adapt to different coverage profiles.

Unlike reference-based approaches, many nodes in a genome graph may remain unaligned due to the presence of more similar alternative paths. As a result, graph-based coverage profiles often contain many zeros and exhibit a sharper signal compared to traditional reference-based coverage.

In summary, tools like `packing` offer an efficient and scalable solution for storing alignment-derived coverage data in graph-based pangenomes. As constructing graphs from

hundreds or thousands of genomes remains computationally intensive, therefore alignment-based variant discovery is likely to remain central for the near future, making efficient coverage storage a critical requirement.

B.3 BVD - Bifurcation variation detection

B.3.1 Introduction

When genome graphs are build with samples from the same species, they represent genetic variation segregating in a population. These pangenomes yield a comprehensive, variation-aware representation that might replace existing traditional linear references in the future. In addition, genome graphs can serve not only as references but also as data structures for direct variation detection, independent of a traditional linear reference. Extracting variation from the graph itself, rather than relative to a reference, is crucial for understanding the genomic dynamics and relationships among the samples used to construct the graph.

These variations are commonly referred to as *bubbles*, and are typically anchored by a unique pair of start and end nodes that define the boundaries of the variant region. In our variation graph, however, these identifiers must include not only the node IDs but also the respective orientations, as directionality is an essential property in bidirected genome graphs.

Bubbles are inherently difficult to identify, as their structures range from simple variants such as SNPs or small indels to more complex forms, including repetitive regions, inversions, or intricate intron–exon arrangements. Depending on the graph construction method, even similar but distantly located genomic sequences may be merged, resulting in large-scale bubbles spanning megabase-sized regions. Moreover, complex bubbles are often non-linear, may involve multiple cycles, and can exhibit internal nesting, depending on the definition applied. These structural complexities make automated detection and interpretation of bubbles a challenging but essential task for understanding genome variation in graph-based representations.

Simplifying path information to single values or bitvectors proved to be ineffective for capturing complex variation structures. To address this limitation, a new algorithm was developed that leverages bifurcation information. This method interprets each variant region as a bifurcation point, where two or more paths (or subpaths from the same sample) diverge within the graph. The algorithm tracks these points of divergence and subsequent reconvergence, identifying the corresponding start and end nodes. The approach operates by iterating over all pairwise path combinations, enabling a comprehensive assessment of the graph's structure and the variation it encodes.

B.3.2 Indexing

The method begins by constructing an index structure that stores the path index for each node. While an initial implementation utilized a `HashMap`, this was later replaced by a

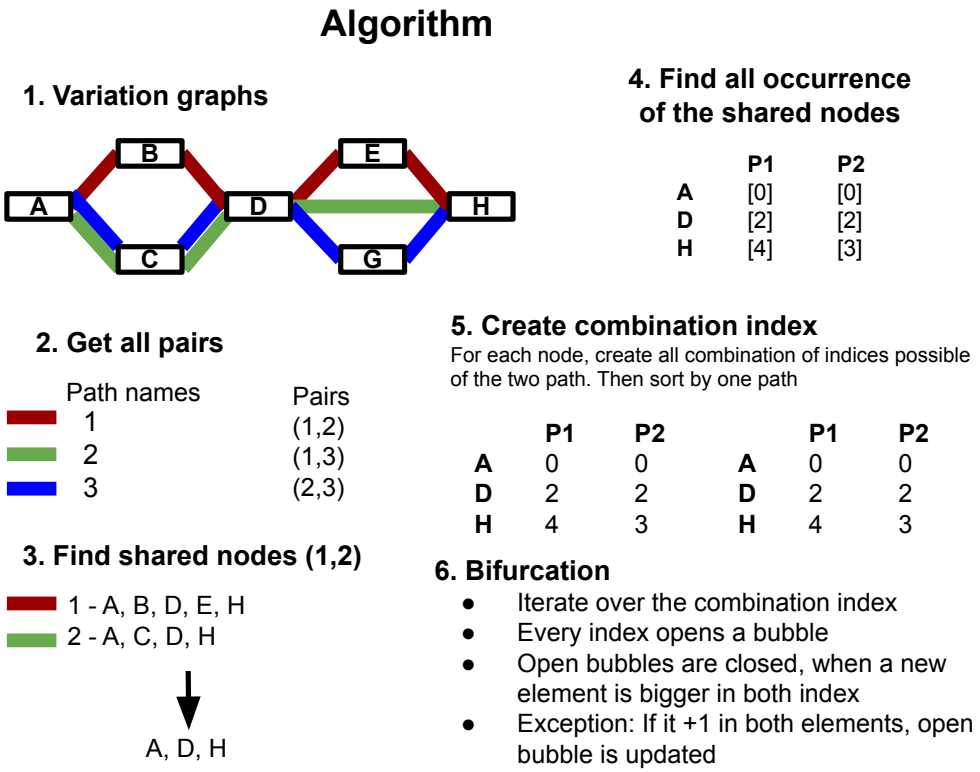


Fig. B.4 Schema of the BVD algorithm.

vector-based lookup table. The improved structure leverages sorted and compacted node identifiers, resulting in enhanced runtime performance and reduced memory usage.

The core of the indexing algorithm involves processing a pair of paths to identify all nodes shared between them. For each shared node, the corresponding path indices are retrieved from the preconstructed lookup table for both paths. All possible combinations of these index pairs are then generated and stored in a vector, hereafter referred to as the combination index or comb-index.

Once the comb-index has been populated for all shared nodes, it is sorted based on one of the path indices. This sorting step facilitates efficient downstream operations, optimizing the process of path comparison and structural analysis within the graph.

B.3.3 Bubble detection

The algorithm operates on the previously constructed comb-vector as input, iterating over each entry sequentially. During this process, a temporary data structure is maintained to store open bubbles, bifurcation that have been initiated but are awaiting closure. At each iteration step, the algorithm evaluates whether the current path indices exceed those of any open bubble. If this condition is satisfied for both paths, the corresponding bubble is complete, and the node identifiers marking the start and end of the bubble are recorded.

A special case arises when both path indices increment by exactly one, indicating that the paths traverse consecutive nodes. In such instances, no divergence occurs, and a new bubble is not formed.

In general, each new index pair is added to the open bubble collection and is either closed or updated in subsequent steps. It is important to note that tips, segments located at the beginning or end of a path that are not shared, are explicitly excluded from bubble reporting 1.

B.3.4 Genomic information and statistics

Although bubble information alone may not always be meaningful, we report genomic intervals (ranges) in the standard BED format. For each interval, we include both the bubble ID and the traversal ID. Traversals refer to the distinct subpaths within a bubble, starting at the initial node and concluding at the terminal node.
In addition, we compile a comprehensive set of statistics for each bubble. This includes the number of traversals and intervals, along with the average, minimum, and maximum lengths of the traversals. We also calculate and report several other features to provide a detailed characterization of each bubble.

B.3.5 Nestedness

Nestedness information is valuable for classifying bubbles based on their structural importance or functional relevance. Although the core algorithm does not inherently support nested relationship detection, these relationships are identified in a post-processing step using path index ranges (intervals). This analysis is performed independently for each path, and the resulting data are subsequently merged.

Within a given path, each interval corresponds to a single bubble. When a bubble occurs multiple times across genomes, it may be represented by multiple intervals. The focus of

this step is to determine the direct nesting relationships, specifically identifying bubbles that are closest to the start and end positions of a given bubble while also fully enclosing it. The objective is to construct a tree-like hierarchy representing the nesting structure among bubbles.

To accomplish this, the algorithm begins by iterating through the sorted set of intervals, discarding those that fall outside the range of interest. For the remaining intervals, it checks whether they encompass the current interval, identifying and recording the nearest enclosing bubble. As disjoint intervals may also exist that partially or fully overlap with the current interval, the process continues until either all candidates have been evaluated or a disjoint enclosing interval is found.

Finally, results from all paths are aggregated into a `HashSet`, producing a unified representation of bubble relationships across the entire graph.

B.3.6 Implementation

The range-based hierarchical overlap detection developed in this work constitutes a standalone component that can be reused in other projects requiring internal nesting logic for overlapping intervals. The approach is built entirely on start and end indices of detected variations and operates independently on each path, making it highly suitable for parallel execution in a multi-threaded environment. The algorithm iterates over sorted ranges and evaluates whether any currently "open" intervals fully enclose the new range. This enables efficient identification of nested structures and supports the construction of tree-like hierarchies within path-specific variation data.

Pseudocode - Bubble detection based on Combination Vector

B.3.7 Parallelism

All processing steps described in this workflow are highly parallelizable, and parallel execution was facilitated using the `rayon` crate in Rust. Path indexing is performed independently for each path, while detection of shared nodes, construction of the comb-index, and bifurcation identification are carried out independently for each pair of paths. Additionally, genomic data extraction and statistical computations are processed in parallel over data chunks. Nestedness relationships can also be precomputed on a per-sample basis, further improving computational efficiency.

Despite these optimizations, empirical observations reveal a high degree of redundancy in the output. For example, when comparing 200 path pairs, it becomes evident that after

Algorithm 1 Bubble detection based on Combination Vector

```
1: function DETECTBUBBLES(input_list)
2:     Input: Sorted index pairs and bubble IDs:
3:         [(i_1, y_1, b_1), (i_2, y_2, b_2), ..., (i_n, y_n, b_n)]
4:     Output: List of bubble ID pairs (b_old, b_new)
5:     open_list ← empty list
6:     bubbles ← empty list
7:     for each new_entry in input_list do
8:         (new_i1, new_i2, new_id) ← new_entry
9:         for each old_entry in temp_list do
10:            (old_i1, old_i2, old_id) ← old_entry
11:            if new_i1 > old_i1 and new_i2 > old_i2 then
12:                if not (new_i1 = old_i1 + 1 and new_i2 = old_i2 + 1) then
13:                    Add (old_id, new_id) to bubbles
14:                end if
15:                                          ▷ Do not add old_entry to updated_temp
16:            end if
17:        end for
18:        Add new_entry to open_list
19:    end for
20:    return bubbles
21: end function
```

approximately 150 comparisons, only minimal novel variation is discovered. Although this redundancy is acknowledged, executing the full pipeline remains essential to ensure comprehensive coverage and to capture the complete spectrum of structural variation present within the dataset.

B.3.8 Complexity

The complexity of our approach primarily depends on the number of paths and nodes involved. We experience quadratic scaling with respect to the number of paths, rendering this method impractical for graphs that contain thousands of paths.

B.3.9 Used data structures

In the initial stages of development, many steps involved the use of hashmaps. However, I restructured all methods and core data structures into a vector-like format. This transition encompassed the index and comb-index structures, as well as an alternative implementation of nodes within our specially developed GFA-reader. The shift to vectors proved advantageous, reducing memory usage, enhancing local data proximity, and consequently improving overall performance.
Our bubble detection algorithm operates exclusively on directed nodes (nodes in combination with their direction). Given that nodes can have two 'outlets' – one positive and one negative – we sought to optimize vector indexing and further minimize memory usage. To achieve this, we combined the direction and node ID into a single u32 integer. This approach allows each integer to represent both attributes using a primitive data type, thereby enhancing the efficiency of our path-index. Nevertheless, we are only allowed to have a maximum count of node IDs of 2.1 billion, which should normally be fine for all present genomes.

B.3.10 Advantages

A major advantages of our algorithm lies in its comprehensive reporting capabilities. We capture and report all types of bubbles present in the graph, ranging from simple small single-nucleotide polymorphisms (SNPs) to more complex structures such as cycles or extensive bubbles that span thousands of base pairs. Additionally, the statistics we generate from this analysis are straightforward and user-friendly. Furthermore, the sequences from the traversals can be easily extracted and compared, thanks to our BED format output, facilitating a more

in-depth and efficient analysis.

B.3.11 Bottleneck

Initially, I had thought that the quadratic complexity was our primary bottleneck, but it turned out to be just a relatively minor issue. With a graph containing 1,011 paths, the workflow requires processing 510,555 pairs. If each run takes 200 ms on a single core, the total computation time amounts to approximately 1.15 days (or 28 hours). While multithreading could potentially reduce this to a matter of minutes, the resources would still be considerable. However, the more significant challenge lay in the creation of the comb-index and the bifurcation algorithm itself. I had initially underestimated the complexity inherent in genome graphs, especially when dealing with highly repeated nodes connected by thousands of different edges. In a benchmarking run with the pan-centromere graph of Chromosome 1, I encountered nodes repeated over 50,000 times. Constructing a comb-index from just two such nodes results in an index size of approximately 1,249,975,000 entries.
A major issue is the prevalence of these highly repeated nodes. They often appear repeatedly in other paths as well, leading to the creation of enormous indexes. This pattern also significantly impacts the generation of genomic intervals, resulting in the creation of millions or even billions of traversals. Calculating relationships becomes impractical with such extensive data.

B.3.12 Optimization strategies

The most substantial performance improvement was achieved by replacing nested `HashMap` structures with vector-based representations. The original double-layered `HashMap` layout was flattened, and each traversal was reported individually within a vector. This change increased memory usage, as both the accession index and bubble identifier had to be stored with each entry. However, the trade-off significantly improved access times and simplified data handling.

Additional optimizations included improving the performance of the GFA parser and introducing multiple layers of multithreading to parallelize I/O and computation more effectively. Looking ahead, further optimizations could involve eliminating the comb-index entirely and performing indexing dynamically during bubble detection. While this would reduce memory consumption, it may introduce a runtime penalty, an important consideration given that execution time is already a primary bottleneck.

A potential optimization strategy to address the mentioned redundancy involves applying stringent filtering to exclude entire ranges of highly repetitive nodes, such as those commonly found in centromeric regions. However, this approach may impose limitations on the types of input graphs that can be processed effectively. In particular, it may reduce generalizability or introduce bias when applied to graphs with varying repeat content or structural complexity.