

Improving Genomic Alignment Accuracy and Variant Detection for Acute Lymphoblastic Leukaemia Patients: A Pan-Genome Graph Approach

Ashlee Thomson-Roberts

ashlee.thomson@sahmri.com

Adelaide University

Jacqueline Rehn

Adelaide University

Barbara McClure

Adelaide University

Laura Eadie

Adelaide University

Elyse Page

Adelaide University

Caitlin Schutz

South Australian Health and Medical Research Institute

Susan Heatley

Adelaide University

David Yeung

Adelaide University

James Breen

The Kids Research Institute Australia

Deborah White

Adelaide University

Research Article

Keywords: Acute lymphoblastic leukaemia, pan-genome graph, variant detection, transcriptomic analysis

Posted Date: September 29th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9098983/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Acute lymphoblastic leukaemia (ALL) is characterised by extensive genomic heterogeneity, including single nucleotide variants, insertions, deletions, and complex rearrangements that are critical for diagnosis, subclassification, and risk stratification. Conventional short-read mRNA sequencing workflows frequently rely on alignment to a single linear reference genome, which incompletely represents disease-associated variation, leading to reference bias, misalignment, and reduced variant detection sensitivity. Pan-genome graphs provide an alternative reference structure by explicitly representing known genomic variation.

Results

We developed a spliced pan-genome graph tailored to acute lymphoblastic leukaemia by integrating disease-associated variants and transcript annotations into the human GRCh38 reference genome. Performance was evaluated using simulated datasets with known ground truth and mRNA sequencing data from ten ALL patients. Graph-based alignment consistently outperformed linear alignment, demonstrating higher mapping precision, improved mapping quality calibration, and increased proportions of Primary and Properly Paired reads. These improvements led to a substantial enhancement in variant detection, as the graph-based approach detected over seven times more unique insertions and deletions and nearly three times more unique single nucleotide variants relative to the linear workflow. Performance gains were most evident for insertions and deletions greater than 10 base pairs, which are commonly overlooked by linear reference alignment approaches when using short-read sequencing data. While graph-based alignment required longer runtime, it achieved lower peak memory usage compared with the linear approach.

Conclusions

A disease-specific spliced pan-genome graph significantly enhances transcriptomic alignment accuracy and variant detection in ALL. The marked improvement in insertion and deletion recovery underscores the clinical value of graph-based references for identifying functionally disruptive variants that are critical for genomic classification. These findings support the adoption of disease-aware pangenome graphs as a practical strategy to improve genomic analyses in precision oncology.

Background

Acute lymphoblastic leukaemia (ALL) is the most common childhood malignancy[1–3] and is characterised by a wide spectrum of genomic alterations[4–6], including single nucleotide variants[7] (SNVs), insertions[8] and deletions[9] (INDELs), and structural rearrangements[10]. Accurate detection of

these events is essential for diagnosis, subtype classification, and treatment stratification[11–14]. However, many clinically relevant variants remain difficult to detect using conventional sequencing approaches, due to their large size, complex structure, or genomic location, underscoring the need for improved genomic analysis methods[6, 15–17].

Standard short-read mRNA sequencing (mRNA-seq) workflows typically rely on mapping reads to a single linear reference genome[18–20]. While this approach has provided valuable insights, it does not fully represent the genetic diversity of the human population or disease-specific alterations[21, 22]. Importantly, reads derived from regions with high variability or structural complexity may be misaligned, leading to reduced sensitivity for variant detection and incomplete capture of the genomic landscape of ALL[15, 23, 24].

Pangenome graphs have emerged as an alternative to linear references by embedding known genomic variation directly into the reference structure[25–27]. In a graph-based framework, multiple alleles and structural variants are represented simultaneously, allowing reads to align more accurately across polymorphic regions[28–30]. Previous studies in both human and non-human systems have demonstrated that graph-based approaches improve alignment accuracy, variant discovery, and the calibration of mapping quality scores[31–33].

Despite these advances, the application of pangenome graphs to cancer remains limited, and disease-specific graph-based references have yet to be extensively explored. ALL is ideal as an early application, with numerous recurrent variants that remain difficult to resolve using linear reference genomes due to their size, location, and complexity[15, 17]. By integrating ALL-associated variants into a spliced pangenome graph, it may be possible to improve both the sensitivity and accuracy of short-read mRNA-seq analysis in this disease context, supporting a scalable approach that takes advantage of the lower cost and greater availability of short-read sequencing compared with long-read methods.

The aim of this study was to construct a disease-specific spliced pangenome graph anchored to the GRCh38 reference genome and enriched with transcript annotations and variants associated with ALL. Anchoring the graph to GRCh38 ensures compatibility with established variant databases and clinical analysis pipelines, facilitating translational applicability. To evaluate the utility of this approach, we benchmarked graph-based mRNA-seq alignment against conventional linear reference alignment using both simulated and real ALL patient datasets. Performance was assessed by evaluating mapping accuracy, variant discovery, and computational efficiency with the goal of evaluating whether disease-specific graph-based references can enhance genomic discovery and provide clinically relevant insights in ALL.

Methods

Simulation of mRNA-seq Reads for Validation

Computationally simulated datasets were generated to enable benchmarking with controlled ground truth datasets. To quantify mapping error rates, simulated mRNA-seq reads were generated. RSEM (v1.3.3)[34] was used to estimate gene and transcript expression levels of an ALL patient, which served as templates for downstream read simulation, ensuring that generated reads reflected the transcriptomic landscape of real ALL patient samples. Using vg sim[21] we simulated 6 data sets. Two different random seed values were used across 3 read depths, 10, 20, and 30 million reads. We used a base substitution error rate of 0.24%[35] and an indel error rate of 0.029%[36]. The alignment tools used in this study employ different mapping quality (MAPQ) scales; the graph-based aligner VG assigns a maximum MAPQ score of 60, whereas the linear mRNA-seq alignment tool STAR[20] uses a maximum of 255. To enable direct comparison, mapping precision distributions were normalised by expressing the proportion of reads assigned the maximum MAPQ score for each aligner as a percentage of total mapped reads.

Real Patient Data and Library Preparation

Simulated sequencing data was complemented by real patient datasets to assess a real-world scenario. These samples were randomly selected. Total RNA was extracted from 10 ALL patient samples (Supplementary Table 1), and mRNA library preparation and paired-end sequencing were conducted as previously described by Thomson et al[19]. Sequencing produced 75-nucleotide paired-end reads, with each sample yielding between 67 and 78 million read pairs. Informed written consent was obtained from all participants, or from a parent or legal guardian for participants under the age of 18 years.

Reference Genome and Variant Database Integration

The GRCh38.p12[37] reference genome was used as the foundational sequence for the pangenome graph construction. Ensembl transcript annotations[38] were integrated into the pangenome structure to create a spliced pangenome graph. This hybrid approach maintains graph connectivity while preserving exon-intron boundaries essential for accurate mRNA-seq transcript mapping.

Variants in the COSMIC database (v0.99)[39] that were listed as being associated with ALL were extracted to create the initial pangenome graph structure. A total of 62,198 exonic variants were integrated into the pangenome graph, representing a comprehensive landscape of ALL-associated genomic alterations (Supplementary Fig. 1). This approach captures both the reference sequence and known ALL-associated variants within a unified graph representation, enabling extensive variant analysis across the ALL-genomic landscape.

Pangenome Graph Construction and Validation

All analyses were performed on a high-performance computing (HPC) cluster comprising seven nodes equipped with Intel(R) Xeon(R) Silver 4309Y CPUs (@2.80 GHz) with 32 threads and 128 GB of memory, running Ubuntu 20.04.5 LTS operating system. Graph construction was performed using the VG toolkit

(v1.55.0)[21]. The reference genome, transcript annotations, and variant files served as inputs to generate a spliced pan-genome graph tailored to ALL-specific features. During graph construction, node identifiers were assigned automatically by VG's internal algorithms, ensuring that every sequence element was uniquely addressable within the graph structure. Construction was performed using a chromosome-by-chromosome approach and then concatenated to optimise memory usage. Graph architecture was assessed using `vg stats`, to calculate the total number of nodes, edges, length in bases and the node ID range.

The completed spliced graph was indexed using VG's `xg` index, which facilitates efficient random access and downstream mapping. Additionally, minimum distance metrics and graph topology information were encoded via VG's distance index, computed using the `vg snarls` utility, which outputs to dedicated files for use in distance-based analyses and variant calling. Path data from the graph were extracted using `vg paths` to enumerate both conventional reference chromosome paths and non-reference sequence paths. Chromosomal path names were filtered using regular expressions to extract canonical chromosome labels.

Testing of the Graph vs Linear reference

Paired-end mRNA-seq reads from both real and simulated datasets were processed using linear and graph-based workflows (Supplementary Figs. 2A and 2B). Raw reads underwent quality control with FastQC (v0.11.9) and MultiQC (v1.8) prior to alignment. For the linear workflow, reads were aligned to the GRCh38 linear reference genome (Linear) using STAR (v2.7.3a), followed by duplicate marking and variant calling with GATK (v4.1.4.1). Variant calls were filtered according to GATK best practices and subsequently separated into SNV and INDEL call sets. For the graph-based workflow, reads were mapped to the spliced pangenome graph using VG (v1.55.0). Graph alignments (GAM) were converted to BAM format using `vg surject`, after which duplicate marking, GATK variant calling, filtering, and SNV/INDEL splitting were performed in an equivalent manner to the linear pipeline to enable direct comparison between approaches.

To evaluate the accuracy and quality of read alignments to both graph-based and linear references, we employed a two-step validation procedure using simulated sequencing data. The graph-based alignments in GAM format were first annotated with mapping information using `vg annotate`, leveraging the spliced graph index. Precision metrics were computed via `vg gamcompare`, which compares the annotated alignments against the known ground truth simulated reads. For linear reference validation, BAM-formatted alignments were converted into GAM format using `vg inject` to ensure compatibility for downstream comparison. Precision scores for linear alignments were then calculated similarly with `vg gamcompare`. These procedures were executed twice per simulated dataset: once with a strict 0 base pair (bp) positional tolerance, reflecting exact positional matches, and once with a relaxed 35 bp tolerance to account for minor positional deviations inherent in alignment processes. This dual-tolerance approach allowed for a nuanced assessment of mapping error rates and alignment goodness scores, capturing both stringent and biologically relevant variations in read placement.

Results

Graph Architecture

Graph architecture describes the structural organisation of a reference graph, including node density, edge connectivity, and branching complexity, all of which directly influence read mapping behaviour, computational performance, and variant resolution. Assessing graph architecture is therefore essential for understanding how the design of a pangenome impacts downstream alignment and variant calling analyses. The ALL-specific pangenome graph (Graph) demonstrated distinct architectural characteristics when compared to existing reference graph implementations[40] (Table 1). Evaluation of graph length, node count, edge count, and edge-to-node ratio[41] across linear, ALL-specific, and population-scale pangenome graphs demonstrated differences in structural complexity, enabling direct assessment of how disease-specific vs population-based variation influences graph organisation and complexity.

Table 1

The graph architecture of a spliced linear reference and the ALL-specific spliced graph, compared to the human pangenome graphs constructed by Liao et al[40]. The Gencode graph was constructed from GRCh38 reference with GENCODE (v38) transcripts. Minigraph used the allele-filtered Minigraph cactus HPRC pangenome with CAT and GENCODE annotations. The 1kg graph incorporated 1000 Genomes Project variants and GENCODE splice junctions.

Build	Length (base pairs)	Number of Nodes	Number of Edges	Node to Edge Ratio	Complexity (vs linear) by node count
Spliced linear	3,088,286,401	97,422,032	97,783,133	1.004	baseline
ALL graph	3,088,362,784	97,570,289	98,004,683	1.004	0.15% increase
Gencode	3,105,715,063	98,036,524	98,425,050	1.004	0.63% increase
Minigraph	3,153,421,135	146,269,282	159,089,525	1.088	50% increase
1kg	3,152,733,168	182,337,610	223,111,989	1.224	87% increase

The linear spliced reference (GRCh38.p12) exhibited the most compact structure, representing the baseline complexity without additional variants. This linear architecture maintained an edge-to-node ratio of approximately 1.004, reflecting the simple sequential connectivity inherent to linear reference structures. As a comparison, the population based human pangenome graphs constructed by Liao et al[40]. demonstrated varying degrees of structural complexity, with their 1kg pangenome graph displaying the highest edge-to-node ratio (1.22) among all evaluated graphs.

Our ALL-specific graph exhibited architectural characteristics closely resembling the linear reference baseline. Despite incorporating 62,198 exonic variants, the ALL graph maintained relatively modest structural complexity, with only a 0.15% increase in nodes and 0.23% increase in edges compared to the linear reference. The total sequence length increased by just 0.002%, indicating that the disease-specific variants were predominantly small-scale alterations rather than large structural variations. The similar

edge-to-node ratios between the ALL graph, Gencode, and linear reference suggest comparable structural organisation, while the higher ratios in Minigraph and 1kg indicate more complex branching patterns typical of population-level variation representation.

This architectural profile demonstrates that the focused incorporation of ALL-associated variants resulted in targeted graph augmentation without the extensive structural complexity observed in population-based pangenome approaches[40].

Mapping Precision Analysis using Simulated Data

A comprehensive evaluation of mapping accuracy was performed on the spliced pangenome graph by using simulated mRNA-seq datasets with known true alignment coordinates. When requiring exact positional accuracy (0 bp tolerance for correct mapping), the graph-based approach demonstrated superior precision compared to linear alignment across all chromosomes and read depths (Fig. 1A, Supplementary Fig. 3A). Alignment to the graph consistently resulted in a lower mapping error rate of approximately 0.001 (99.9% precision) across all chromosomes, regardless of read depth. In stark contrast, linear alignment showed significantly higher and more variable mapping errors, ranging from 0.003 to 0.010 (97.0% to 99.7% precision) across different chromosomes.

The most pronounced differences were observed on chromosomes 4 and 12, where alignment to the linear reference resulted in mapping errors of ~ 0.010 (99.0% precision) while the graph-based reference maintained a consistent error rate of ~ 0.001 (99.9% precision). This represents a 10-fold improvement in mapping accuracy for these chromosomal regions. Notably, both approaches showed remarkable consistency across read depths, with 10, 20, and 30 million read datasets producing nearly identical precision values, indicating robust scalability of both the graph- and linear-based methods.

Even when allowing a more permissive 35 bp tolerance for mapping accuracy, the pangenome graph approach maintained its precision advantage over linear alignment (Fig. 1B, Supplementary Fig. 3B). Under these relaxed criteria, the graph achieved mapping errors consistently around 0.0002 across all chromosomes (99.98% precision), while the linear showed improved but still inferior performance with errors ranging from 0.0008 to 0.001 (99.9% to 99.92% precision). This sustained advantage under relaxed mapping criteria demonstrates that the pangenome graph approach provides superior accuracy not only for exact read placement but also for approximate genomic localisation. The 4- to 5-fold improvement in precision under permissive criteria indicates that graph-based alignment fundamentally improves read placement accuracy rather than simply achieving marginal gains through strict positioning.

Mapping Quality Calibration Assessment using Simulated Data

To evaluate the quality calibration of mapping assignments, we analysed mapping goodness scores, which quantify both overall accuracy and the reliability of MAPQ score assignments. The mapping

goodness score rewards aligners that assign high MAPQ scores only to genuinely correct alignments while penalising situations where reads with high MAPQ scores are frequently incorrect. Analysis of mapping goodness scores using a stringent 0 bp tolerance revealed that alignment to the graph-based reference consistently achieved superior quality calibration compared to the linear across all chromosomes and read depths (Fig. 2A, Supplementary Fig. 4A). The graph demonstrated mapping goodness scores ranging from 1.8 to 3.2, with generally higher values indicating better calibration between assigned confidence scores and actual mapping accuracy.

Under the more relaxed 35 bp tolerance criteria (Fig. 2B, Supplementary Fig. 4B), the graph maintained superior mapping goodness scores (ranging from 2.0 to 6.0) compared to the linear. The sustained advantage in quality calibration across different stringency levels demonstrates that the pangenome graph approach provides not only more accurate mappings but also more reliable confidence estimates for those mappings.

The enhanced quality calibration observed in our graph approach demonstrates that we can more reliably employ MAPQ scores for downstream filtering and analysis, as high-confidence assignments derived from the graph-based approach are more likely to reflect truly accurate mappings.

Alignment Assessment using Real Patient Data

Since the true origin of sequencing reads in patient-derived data is unknown, mapping accuracy cannot be assessed directly. Therefore, Primary alignment metrics are used as an indirect indicator of alignment performance. The comparison of Primary alignments between graph-based and linear reference mapping reveals a statistically significant difference ($p < 0.05$), with graph-based mapping achieving consistently higher percentages of Primary alignments (Fig. 3A). The distribution analysis shows that linear mapping typically achieves Primary alignment rates clustering around 96–97% of total reads, while graph-based mapping demonstrates superior performance with Primary alignment rates predominantly in the 99–99.5% range. This improvement in Primary alignment rates represents a substantial enhancement in data utilisation, particularly critical in ALL research where capturing the full spectrum of genomic alterations is essential for accurate diagnosis and treatment stratification. The tight distribution of graph-based mapping results also indicates greater consistency and reliability across different patient samples, reducing inter-sample variability that could confound downstream analyses.

The analysis of Properly Paired alignments (Fig. 3B) provides additional validation of pangenome superiority in handling the complex genomic landscape of ALL. There was a significant increase ($p \leq 0.005$) in Properly Paired reads when using the graph-based approach. This improvement in Properly Paired alignments is particularly crucial for ALL analysis, as proper read pairing is essential for accurate detection of structural variants, including gene fusions and complex rearrangements that are hallmarks of this disease. The enhanced ability to maintain read pair integrity suggests that pangenome graphs better accommodate the diverse genomic architectures present in ALL patient samples, leading to more reliable variant calling and improved detection of clinically relevant genomic alterations.

Variant Detection Performance Assessment using Real Patient Samples

To evaluate the practical utility of graph-based alignment for variant discovery, we compared its performance against a traditional linear reference approach using mRNA-seq data from real ALL patient samples ($n = 10$) (Supplementary Table 1). While many variants detected were single nucleotide variants (SNVs), both the graph and linear approaches showed similar distributions of reported variant sizes. (Fig. 4A, Supplementary Fig. 5A-J). However, the graph-based approach consistently detected higher numbers of variants across most size categories compared to the linear, particularly for deletions and insertions larger than 10bp. This enhanced detection capability suggests that the graph structure enables more sensitive identification of large insertions and deletions that may be missed by linear alignment approaches.

Analysis of SNV detection (Fig. 4B, Supplementary Fig. 6A-J) showed that the graph and the linear achieved substantial overlap in variant calling, indicating good overall concordance between approaches. However, the pangenome graph approach demonstrated superior sensitivity, detecting an additional 1,587,982 unique SNVs (26.6% of total detected variants) that were not reported when using the linear reference approach. This represents a 41.5% increase in SNV detection capability compared to the variants uniquely detected by the linear (555,345 SNVs, 9.3% of total). The 2.86-fold higher number of unique SNVs detected by the graph compared to the linear indicates that pangenome graph alignment provides access to previously hidden genetic variation, which is particularly valuable for understanding the full spectrum of genomic alterations in ALL patients. Analysis of INDEL detection revealed similar patterns of superior sensitivity for the pangenome graph approach (Fig. 4C, Supplementary Fig. 6A-J). The graph detected 421,142 unique INDELs (46.7% of total INDELs detected) compared to 56,779 unique INDELs (6.3%) detected exclusively by the linear, representing a 7.42-fold improvement in unique INDEL discovery.

Computational Performance Analysis

Comparative analysis of computational runtime revealed expected trade-offs between accuracy and processing speed (Fig. 5A). Linear alignment demonstrated superior processing speed, achieving an average of 3.5 million reads mapped per minute across the 10 patient samples (range: 2.4–3.8 million reads/minute). In contrast, the graph-based alignment averaged 0.47 million reads mapped per minute (range: 0.3–0.5 million reads/minute), representing a 7.4-fold difference in processing speed

Analysis of peak memory usage revealed a surprising advantage for the pangenome graph approach (Fig. 5B). The graph-based alignment required an average of 27.6 GB of peak memory (range: 27.5–28.5 GB) compared to 36.5 GB for the linear alignment across all patient samples. This represents a 24.4% reduction in memory requirements for the graph-based approach. The lower memory footprint of the graph is particularly notable given the additional complexity of the graph data structure and the superior variant detection performance.

Discussion

In this study, we showed the advantage of a disease-specific pangenome graph over linear reference alignment for transcriptomic analysis in ALL, using simulated datasets, then validated in a real patient dataset. We found superior performance using the pangenome graph approach, as assessed on mapping accuracy, computational performance, mapping quality calibration, and downstream variant detection. Given the central role of genomic alterations in defining ALL subtypes[2, 6, 42, 43], such improvements have direct clinical relevance.

Compared to the human pangenome graphs[40, 44], our disease-specific approach prioritises clinically relevant variation specific to ALL. This targeted strategy results in a graph that is structurally simpler yet functionally optimised for disease context, improving performance for ALL-specific analyses while avoiding the computational overhead associated with resolving population variations of lesser significance. This approach improved performance, while keeping computational complexity to a scale comparable to linear alignments. Regardless, the computational performance analysis reveals expected trade-offs between accuracy and processing speed. The graph alignment demonstrated higher throughput, with increased resources. It is important to contextualise this difference within the broader clinical workflow and weigh it against the substantial improvements in accuracy and variant detection. For routine clinical diagnostics where turnaround time is critical, the additional computational time required for graph alignment represents a significant consideration. However, several factors may mitigate this concern. The computational cost is partially offset by the graph's 24% reduction in peak memory usage which may enable parallel processing of multiple samples on shared computational resources. However, this memory efficiency may reflect differences in the internal algorithms and data structures used by the respective alignment tools (VG vs STAR) rather than fundamental properties of graph vs linear alignment. As graph-based tools continue to mature and optimise, further runtime improvements can be anticipated.

Using simulated data, pangenome graph-based alignment substantially outperforms linear reference mapping in terms of positional accuracy, regardless of genomic location and read depth. A variant allele can map precisely to the corresponding alternative path in a graph, whereas in linear alignment, mapping will incur a penalty in mismatching due to reference bias[45, 46]. The pangenome graph, under various conditions, improves fundamental read placement accuracy rather than merely achieving marginal gains through strict positioning. The improved precision is intrinsic to the alignment methodology rather than artifacts of sampling depth or statistical power and likely ease its implementation in clinical settings where sequencing depth may vary across samples due to technical or economic constraints. Validated by real patient data, the improvement in Primary and Properly Paired alignment rates is particularly significant for structural variant detection and transcript isoform analysis. Proper read pairing is essential for accurate detection of fusion transcripts, large deletions, and complex rearrangements, genomic features that are hallmarks of ALL pathogenesis and critical determinants of disease subtype and prognosis. The graph's superior maintenance of read pair integrity, suggests that it better accommodates the diverse genomic architectures present in ALL patient samples, including regions

with complex variation that may disrupt read pairing in linear alignment. Furthermore, the tight distribution and consistency of graph-based alignment metrics across patient samples indicates robust performance independent of sample-specific characteristics such as variant burden, transcript expression profiles, or sequencing quality.

The dramatic improvement in INDEL detection (7.42-fold) compared to SNV detection (2.86-fold) likely reflects the greater challenges that INDELs pose for linear alignment. INDELs can disrupt read alignment more severely than SNVs, potentially causing reads to be discarded or misaligned when using a linear reference. The graph's ability to represent known INDELs as alternative paths allows reads spanning these variants to align correctly, maintaining coverage and enabling detection of additional nearby variants. For ALL clinical applications, this enhanced variant detection capability is particularly valuable given that INDELs are often more disruptive to protein function than SNVs, and may reveal previously unrecognised disease-relevant alterations, potentially improving diagnostic accuracy.

This dramatic improvement in variant detection sensitivity can be attributed to multiple complementary mechanisms. First, the graph's ability to accurately map reads containing known variants prevents these reads from being discarded or misaligned, thereby maintaining coverage across variant-containing regions and enabling detection of additional variants in proximity to known alterations. Second, the improved mapping accuracy and MAPQ calibration result in higher-quality alignments that meet variant caller confidence thresholds, reducing false negative rates. The substantial overlap between graph and linear variant calls demonstrates that the graph approach is not simply detecting different variants but rather provides a superset of the variants detectable by linear methods. This suggests that adopting graph-based alignment would not require abandoning established variant interpretation frameworks, rather it would strengthen existing detection capabilities with access to previously hidden genomic variation.

Beyond raw mapping accuracy, improved mapping goodness scores with pangenome graph provides superior calibration of mapping quality (MAPQ) scores. This is critically important for downstream variant calling and filtering workflows that rely on MAPQ thresholds to distinguish high-confidence variant calls from potential false positives. The superior mapping goodness scores achieved by the graph indicate that when the graph-based aligner assigns a high MAPQ score, that assignment is more reliably indicative of correct alignment. Conversely, the graph's ability to represent alternative alleles allows it to more accurately identify and assign low confidence scores to genuinely ambiguous mappings, where a read could plausibly originate from multiple genomic locations. This improved calibration likely stems from the graph's explicit representation of known variants, which provides the alignment algorithm with richer contextual information for evaluating mapping ambiguity. In linear alignment, a read containing a variant allele may appear to map poorly to the reference and thus receive a low MAPQ score even if it maps to the correct genomic location. In contrast, the graph can recognise that the read matches a known variant path perfectly, justifying assignment of a high MAPQ score. Improved MAPQ calibration translates directly to more reliable variant filtering and reduced false discovery rates. Variant callers can apply standard MAPQ thresholds with greater confidence that

retained variants represent true biological variation rather than alignment artifacts, potentially reducing the burden of manual curation and improving the efficiency of clinical genomic analysis workflows.

While our study demonstrates substantial advantages of disease-specific pangenome graphs, several limitations warrant consideration. Our graph was constructed using variants from the COSMIC database[39], which primarily captures known somatic alterations identified in previous studies. This approach may not represent the full spectrum of ALL-associated variation, particularly rare or novel variants that have not yet been catalogued. Future iterations could benefit from iterative graph refinement incorporating variants discovered through graph-based analysis itself, creating a progressively more comprehensive representation of ALL genomic diversity.

Our evaluation focused on mRNA-seq data, which captures only the transcribed portion of the genome. While this is appropriate for the analysis of many ALL samples, where many clinically relevant alterations affect expressed genes and transcripts, it limits our ability to assess graph performance for non-coding variants that may not be detectable in transcriptomic data. Extension of this approach to whole-genome or whole-exome sequencing data would provide a more comprehensive evaluation of graph utility across the full genomic landscape, although it would also lead to increased complexity.

The simulated data used for precision evaluation, while valuable for providing ground truth alignments, may not fully recapture the complexity and error profiles of real sequencing data. Although we incorporated realistic error rates based on published sequencing characteristics, factors such as batch effects, sequence-specific biases, and library preparation artifacts may influence real-world performance in ways not captured by simulation. Our validation using real patient data partially addresses this limitation, though without ground truth alignments, we cannot directly quantify mapping accuracy in these datasets. Our variant calling analysis compared total variant counts between graph and linear approaches but did not include orthogonal validation to definitively establish which approach yields more true positive variants versus false positives. While the improved mapping accuracy observed in simulated data strongly suggests that the additional variants detected by the graph represent true biological variation, experimental validation through alternative sequencing technologies (e.g., long-read sequencing or targeted validation) would provide definitive confirmation.

An additional limitation of the current study is that not all structural variants (SVs) were explicitly incorporated into the pangenome graph. Structural alterations such as large duplications, inversions, and complex rearrangements are a defining feature of ALL pathogenesis, underpinning recurrent gene fusions, copy number changes, and chromosomal translocations that drive disease biology. Future improvements will incorporate ALL-specific structural variants to further enhance alignment accuracy, fusion detection, and breakpoint resolution, particularly for clinically actionable rearrangements.

Conclusion

In this study we demonstrate that a spliced pangenome graph built on GRCh38 and enriched with ALL-associated variants substantially improves mRNA-seq read alignment, mapping quality calibration, and

discovery sensitivity of both SNVs and INDELs when compared with conventional linear reference alignment. Across simulated benchmarks and ALL patient samples, the graph-based approach delivered consistently higher mapping precision (at both exact and relaxed tolerances), superior MAPQ calibration, increased proportions of Primary and Properly Paired alignments, and lower peak memory usage, albeit at a substantial runtime cost. The most significant advantage, however, was observed in variant detection, where recovery of unique mid-to-large indels was markedly improved. Together, these results indicate that incorporating disease-specific variation into reference structures can considerably increase the biological information recovered from mRNA-seq data and thus has clear potential to improve genomic characterisation in disease specific settings.

Declarations

Ethics approval and consent to participate

Ethical approval for sample collection and analysis was obtained from the Central Adelaide Local Health Network (CALHN) Human Research Ethics Committee and the Women's and Children's Health Network (WCHN) Human Research Ethics Committee. Approval for the collection of samples outside of South Australia was granted under the National Mutual Acceptance Scheme. Informed written consent was obtained from all participants, or from a parent or legal guardian for participants under the age of 18 years. All experiments were performed in accordance with the principles outlined in the Declaration of Helsinki.

Consent for publication

Not Applicable.

Competing interests

The authors declare no competing interests.

Funding

This work was in part supported by funding from the Australian Genomics Health Alliance (AGHA) (1113531 to D.L.W.), Medical Research Future Fund (MRFF) (2007441 to D.L.W.), Beat Cancer (Project grant x2 & Translational Research Package to D.L.W., www.cancersa.org.au), and the Leukaemia Foundation (D.L.W., www.leukaemia.org.au). D.L.W. is an NHMRC R.D. Wright II Fellow (1160833) and a Beat Cancer Principal Research Fellow. L.N.E is supported by the Cancer Council SA's Peter Nelson Leukaemia Research Fellowship (www.cancersa.org.au) on behalf of its donors; and the Cancer Council SA's Beat Cancer Project on behalf of its donors and the State Government through the Department of Health D.T.Y. holds an Early-Career Clinician Researcher Fellowship from South Australia Cancer Council.

The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contribution

Conceptualisation, A.J.T.; Data curation, A.J.T. and J.A.R.; Formal analysis, A.J.T.; Funding acquisition, D.T.Y. and D.L.W.; Investigation, S.L.H., L.N.E., E.C.P., C.S., B.J.M.; Methodology, A.J.T. ; Project administration, J.A.R.; Resources, D.L.W.; Software, A.J.T.; Supervision, J.A.R., D.T.Y., J.B. and D.L.W.; Visualisation, A.J.T.; Writing - original draft, A.J.T.; Writing - review & editing, J.A.R., S.L.H., L.N.E., E.C.P., C.S., B.J.M., D.T.Y., J.B. and D.L.W. All authors have read and agreed to the published version of the manuscript.

Acknowledgement

We would like to acknowledge the ongoing support of SAHMRI, NHMRC, The Cancer Council of South Australia and The Leukaemia Foundation Australia, as well as the South Australian Genomics Centre (SAGC, SA, AUS) for their sequencing services. Finally, we would like to thank the patients, their families and treating clinicians for their participation in this study.

Data Availability

The datasets generated and/or analysed during the current study (EGAS50000001807) are available in the European Genome-phenome Archive (EGA) repository (EGAD50000002606). All data generated or analysed during this study are included in this published article and its supplementary information files. Code and additional scripts needed to reproduce the method seen here can be found in the Git repository (https://github.com/ashleethomson/ALL_GenomeGraph).

References

1. Duffield AS, Mullighan CG, Borowitz MJ. International Consensus Classification of acute lymphoblastic leukemia/lymphoma. *Virchows Arch*. 2022. <https://doi.org/10.1007/s00428-022-03448-8>.
2. Elgazar S, Constantinou C. Paediatric acute lymphoblastic leukaemia: A narrative review of current knowledge and advancements. *Curr Oncol Rep*. 2024;:1–14.
3. Ma SB, Lin W, Campbell J, Clerici K, White D, Yeung D, et al. Laboratory validation and clinical utility of next-generation sequencing-based IGH/TCR clonality testing for the monitoring of measurable residual disease in acute lymphoblastic leukaemia: real-world experience at Austin Pathology. *Pathology*. 2024. <https://doi.org/10.1016/j.pathol.2024.04.012>.

4. Vicente-Garcés C, Fernández G, Esperanza-Cebollada E, Richarte-Franqués M, Crespo-Carrasco A, Montesdeoca S, et al. RNA-sequencing: A reliable tool to unveil transcriptional landscape of paediatric B-other acute lymphoblastic leukaemia. *Br J Haematol*. 2025. <https://doi.org/10.1111/bjh.20056>.
5. Vllahu M, Savarese M, Cantiello I, Munno C, Sarcina R, Stellato P, et al. Application of omics analyses in pediatric B-cell acute lymphoblastic leukemia. *Biomedicines*. 2025;13:424.
6. Pölönen P, Di Giacomo D, Seffernick AE, Elsayed A, Kimura S, Benini F et al. The genomic basis of childhood T-lineage acute lymphoblastic leukaemia. *Nature*. 2024;:1–10.
7. Gu Z, Churchman ML, Roberts KG, Moore I, Zhou X, Nakitandwe J, et al. PAX5-driven subtypes of B-progenitor acute lymphoblastic leukemia. *Nat Genet*. 2019;51:296–307.
8. Li Z, Chang T-C, Junco JJ, Devidas M, Li Y, Yang W, et al. Genomic landscape of Down syndrome-associated acute lymphoblastic leukemia. *Blood*. 2023;142:172–84.
9. Kobitzsch B, Gökbüget N, Schwartz S, Reinhardt R, Brüggemann M, Viardot A, et al. Loss-of-function but not dominant-negative intragenic IKZF1 deletions are associated with an adverse prognosis in adult BCR-ABL-negative acute lymphoblastic leukemia. *Haematologica*. 2017;102:1739–47.
10. Israeli S. Deciphering ‘B-others’: Novel fusion genes driving B-cell acute lymphoblastic leukemia. *EBioMedicine*. 2016;8:8–9.
11. Tran TH, Hunger SP. The genomic landscape of pediatric acute lymphoblastic leukemia and precision medicine opportunities. *Semin Cancer Biol*. 2022;84:144–52.
12. Li J-F, Dai Y-T, Lilljebjörn H, Shen S-H, Cui B-W, Bai L, et al. Transcriptional landscape of B cell precursor acute lymphoblastic leukemia based on an international study of 1,223 cases. *Proc Natl Acad Sci U S A*. 2018;115:E11711–20.
13. Hamadeh L, Enshaei A, Schwab C, Alonso CN, Attarbaschi A, Barbany G, et al. Validation of the United Kingdom copy-number alteration classifier in 3239 children with B-cell precursor ALL. *Blood Adv*. 2019;3:148–57.
14. Yeung DT, Eadie LN, Rehn J, Heatley SL, McClure BJ, Page EC et al. Diagnostic genomic analysis is prognostic in AYA ALL patients treated on a MRD-stratified pediatric protocol. *Blood Neoplasia*. 2024;:100041.
15. Thomson A, Rehn J, Yeung D, Breen J, White D. Deciphering IGH rearrangement complexity and detection strategies in acute lymphoblastic leukaemia. *NPJ Precis Oncol*. 2025;9:99.
16. Eastmond DA, Schuler M, Rupa DS. Advantages and limitations of using fluorescence in situ hybridization for the detection of aneuploidy in interphase human cells. *Mutat Res*. 1995;348:153–62.
17. Rehn JA, O’Connor MJ, White DL, Yeung DT. DUX Hunting-Clinical Features and Diagnostic Challenges Associated with DUX4-Rearranged Leukaemia. *Cancers*. 2020;12. <https://doi.org/10.3390/cancers12102815>.
18. Haas BJ, Dobin A, Li B, Stransky N, Pochet N, Regev A. Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biol*.

- 2019;20:1–16.
19. Thomson AJ, Rehn JA, Heatley SL, Eadie LN, Page EC, Schutz C, et al. Reproducible Bioinformatics Analysis Workflows for Detecting IGH Gene Fusions in B-Cell Acute Lymphoblastic Leukaemia Patients. *Cancers*. 2023;15:4731.
 20. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29:15–21.
 21. Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat Biotechnol*. 2018;36:875–81.
 22. Jiao D, Dong X, Fan S, Liu X, Yu Y, Wei C. Gastric cancer genomics study using reference human pangenomes. *Life Sci Alliance*. 2025;8. <https://doi.org/10.26508/lsa.202402977>.
 23. Bhagwat AS, Lu B, Vakoc CR. Enhancer dysfunction in leukemia. *Blood*. 2018;131:1795–804.
 24. Ratti S, Lonetti A, Follo MY, Paganelli F, Martelli AM, Chiarini F, et al. B-ALL Complexity: Is Targeted Therapy Still A Valuable Approach for Pediatric Patients? *Cancers*. 2020;12. <https://doi.org/10.3390/cancers12123498>.
 25. Tao Y, Jordan DR, Mace ES. A Graph-Based Pan-Genome Guides Biological Discovery. *Mol Plant*. 2020;13:1247–9.
 26. Secomandi S, Gallo GR, Rossi R, Rodríguez Fernandes C, Jarvis ED, Bonisoli-Alquati A et al. Pangenome graphs and their applications in biodiversity genomics. *Nat Genet*. 2025;:1–14.
 27. Chikhi R, Dufresne Y, Medvedev P. Constructing and personalizing population pangenome graphs. *Nat Methods*. 2024;:1–2.
 28. Sanaullah A, Villalobos S, Zhi D, Zhang S. Haplotype Matching with GBWT for Pangenome Graphs. *bioRxiv*. 2025. <https://doi.org/10.1101/2025.02.03.634410>
 29. Du Z-Z, He J-B, Jiao W-B. A comprehensive benchmark of graph-based genetic variant genotyping algorithms on plant genomes for creating an accurate ensemble pipeline. *Genome Biol*. 2024;25:91.
 30. Levy-Sakin M, Pastor S, Mostovoy Y, Li L, Leung AKY, McCaffrey J, et al. Genome maps across 26 human populations reveal population-specific patterns of structural variation. *Nat Commun*. 2019;10:1025.
 31. Leonard AS, Crysnanto D, Mapel XM, Bhati M, Pausch H. Graph construction method impacts variation representation and analyses in a bovine super-pangenome. *Genome Biol*. 2023;24:124.
 32. Rice ES, Alberdi A, Alfieri J, Athrey G, Balacco JR, Bardou P, et al. A pangenome graph reference of 30 chicken genomes allows genotyping of large and complex structural variants. *BMC Biol*. 2023;21:267.
 33. Du Z-Z, He J-B, Jiao W-B. Plant graph-based pangenomics: techniques, applications, and challenges. *Abiotech*. 2025;:1–16.
 34. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. 2011;12:323.

35. Pfeiffer F, Gröber C, Blank M, Händler K, Beyer M, Schultze JL, et al. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Sci Rep*. 2018;8:10950.
36. Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol*. 2018;36:983–7.
37. Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen H-C, Kitts PA, et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *bioRxiv*. 2016;072116. <https://doi.org/10.1101/072116>.
38. Dyer SC, Austine-Orimoloye O, Azov AG, Barba M, Barnes I, Barrera-Enriquez VP, et al. Ensembl 2025. *Nucleic Acids Res*. 2025;53:D948–57.
39. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res*. 2019;47:D941–7.
40. Liao W-W, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, et al. Draft Hum Pangenome Ref *bioRxiv*. 2022. <https://doi.org/10.1101/2022.07.09.499321>. :2022.07.09.499321.
41. Beyer W, Novak AM, Hickey G, Chan J, Tan V, Paten B, et al. Sequence tube maps: making graph genomes intuitive to commuters. *Bioinformatics*. 2019;35:5318–20.
42. Krali O, Marincevic-Zuniga Y, Arvidsson G, Enblad AP, Lundmark A, Sayyab S, et al. Multimodal classification of molecular subtypes in pediatric acute lymphoblastic leukemia. *NPJ Precis Oncol*. 2023;7:131.
43. Zhang L, Habeebu SSM, Li W. Prognostic and Predictive Biomarkers in Precursor B-cell Acute Lymphoblastic Leukemia. *Exon*; 2022.
44. Liao W-W. Construction of a Human Pangenome Reference to Improve Structural Variation Detection. 2023.
45. Paten B, Novak AM, Eizenga JM, Garrison E. Genome graphs and the evolution of genome inference. *Genome Res*. 2017;27:665–76.
46. Martiniano R, Garrison E, Jones ER, Manica A, Durbin R. Removing reference bias and improving indel calling in ancient DNA data analysis by mapping to a sequence variation graph. *Genome Biol*. 2020;21:250.

Figures

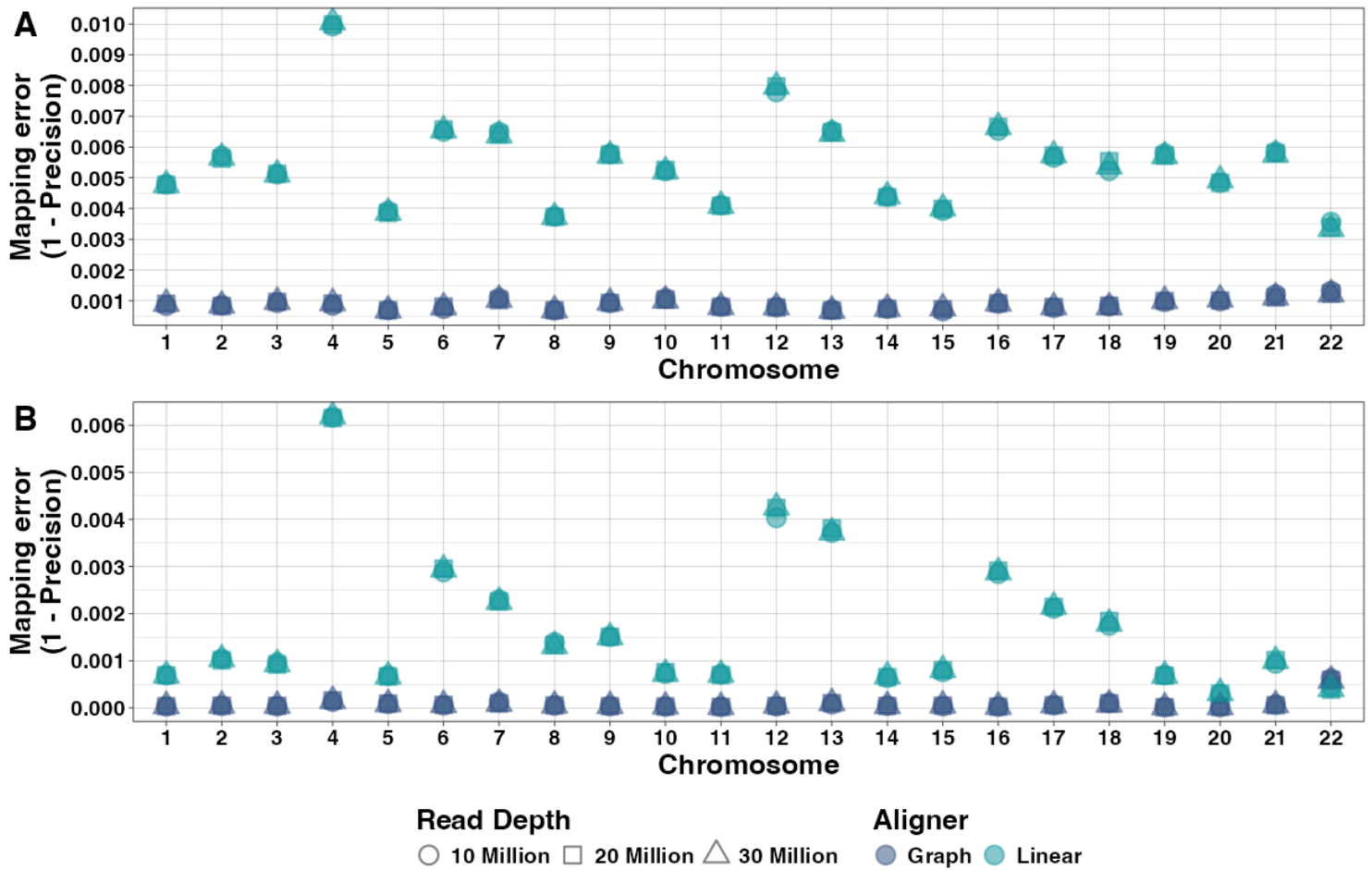


Figure 1

Mapping error rates across chromosomes for one of the simulated data sets under two tolerance settings. Mapping error rates are shown for each chromosome using graph-based and linear alignment at three read depths. **(A)** Error rates with a 0 bp positional tolerance. **(B)** Error rates with a 35 bp tolerance. Both panels use the same simulated dataset.

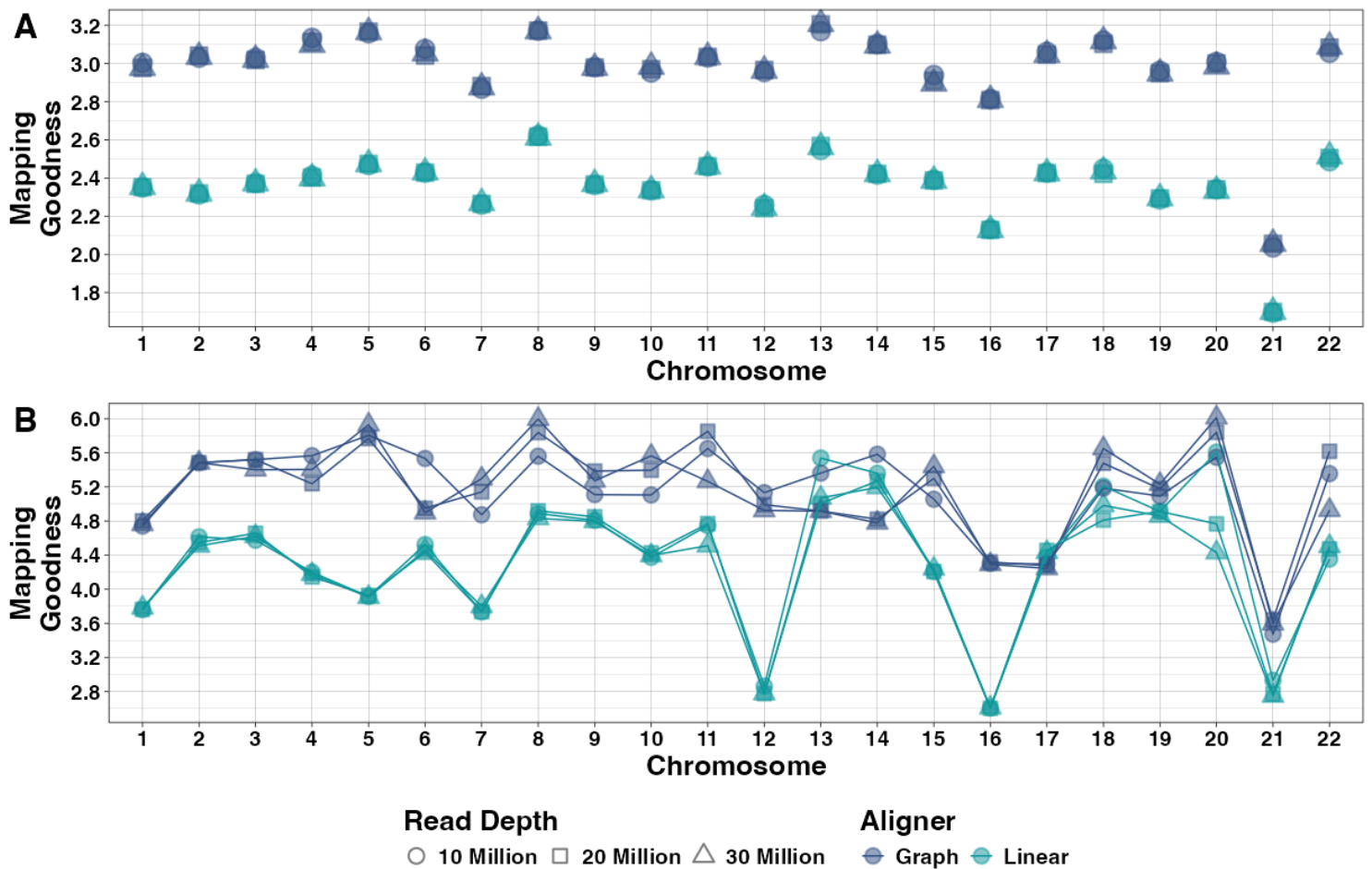


Figure 2

Mapping goodness scores across chromosomes for one of the simulated data sets under two tolerance settings. Mapping goodness scores are shown for each chromosome for graph-based and linear alignment at three read depths. **(A)** Scores calculated with a 0 bp positional tolerance. **(B)** Scores calculated with a 35 bp tolerance. Both panels use the same simulated dataset.

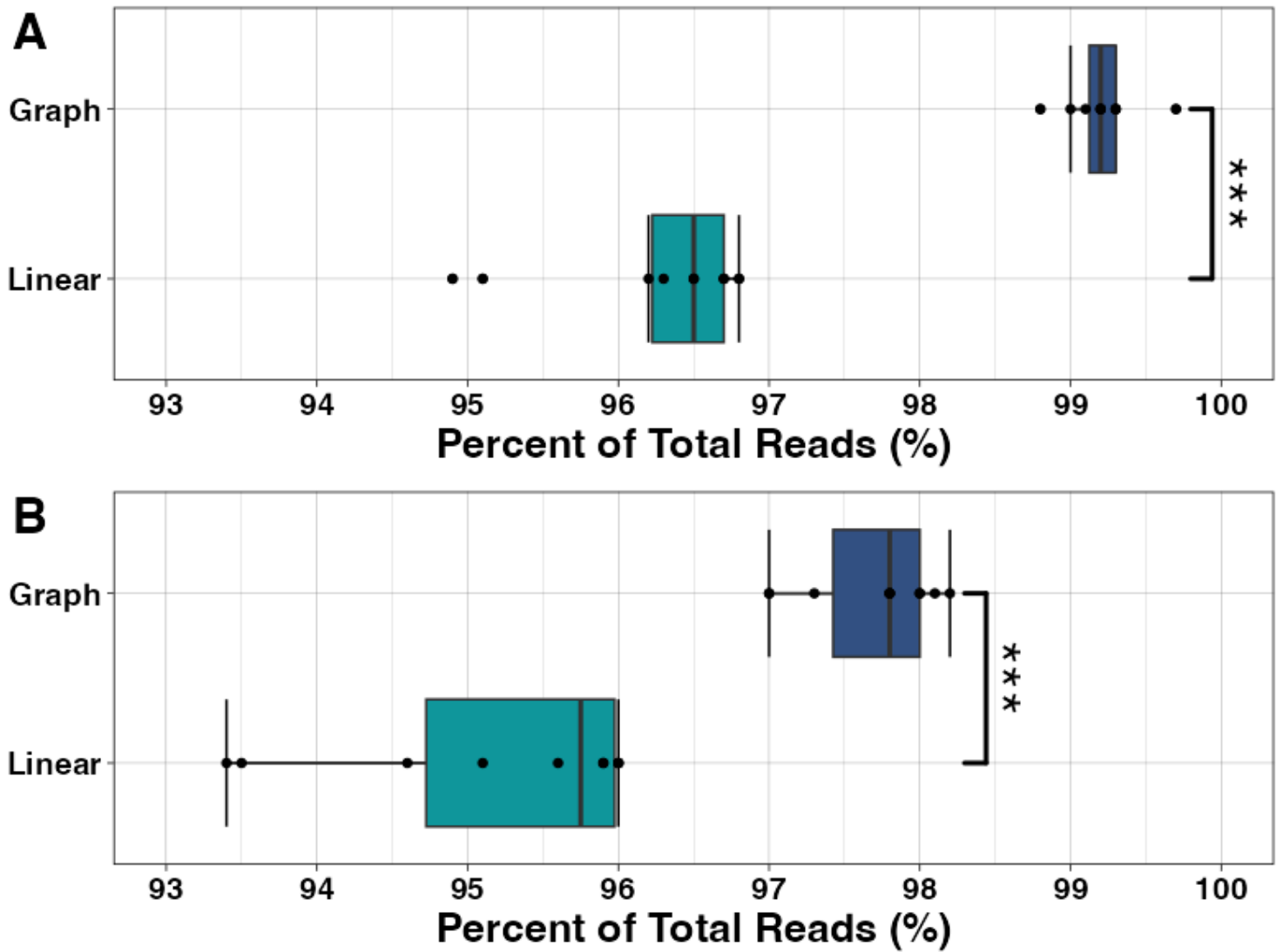


Figure 3

Percent of total reads mapped by graph-based and linear alignment for real patient data sets. Comparison of sequencing reads mapped to linear vs graph was determined using Paired t-test in R, where p-values of $p \leq 0.005$ (***) were considered statistically significant. The graph-based mapping achieved a statistically significant higher proportion of mapped reads than linear alignment for both the **(A)** Primary alignments ($p = 4.3 \times 10^{-7}$), and **(B)** the Properly Paired alignments ($p = 1.5 \times 10^{-5}$).

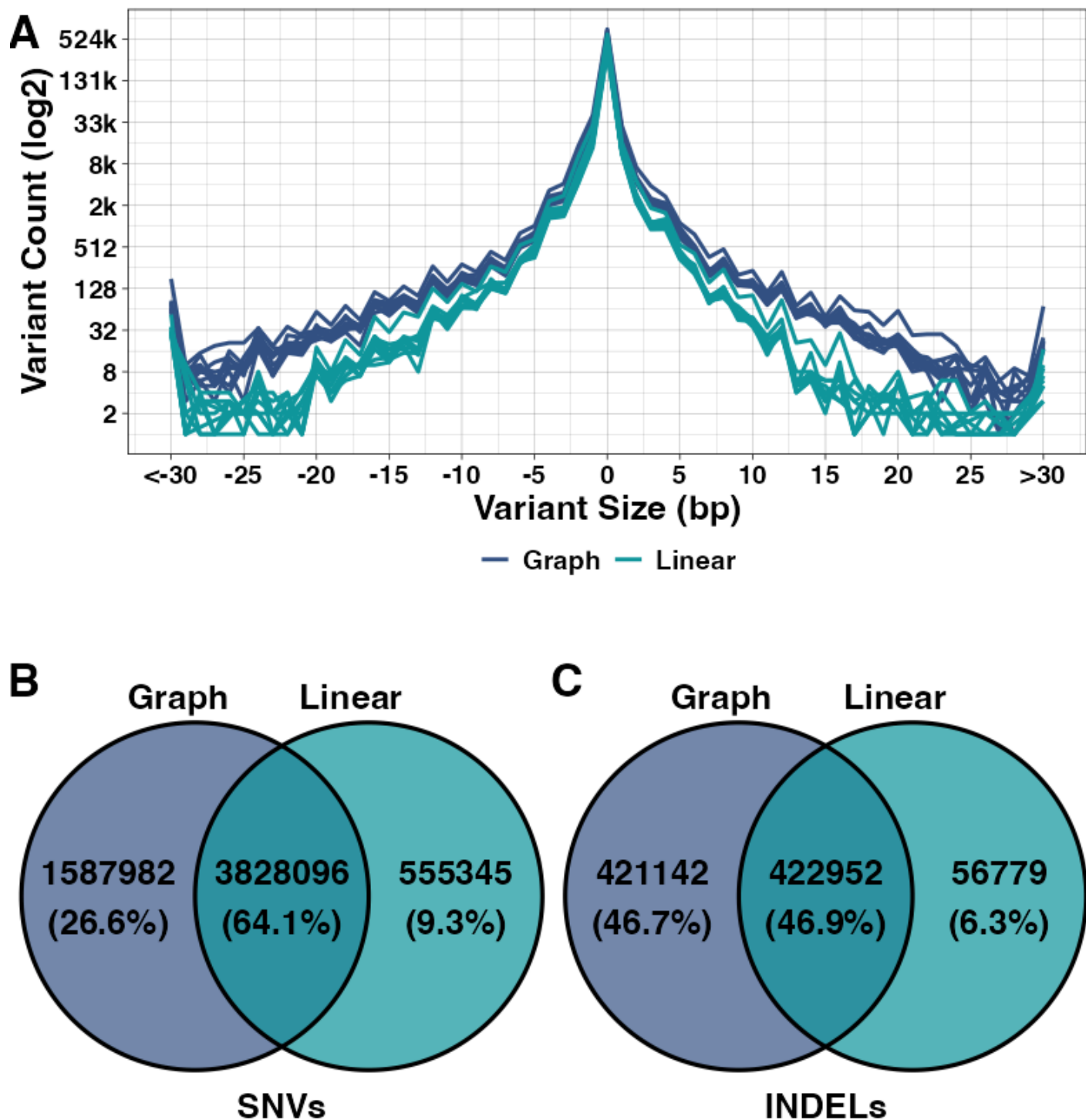


Figure 4

Comparison of variant calls generated using the graph-based approach vs linear reference. **(A)** Distribution of insertion and deletion sizes identified across the 10 real patient samples. Variant counts (log₂ scale) are plotted by size for both graph-based and linear alignments, showing differences in detection sensitivity across the full INDEL spectrum. **(B, C)** Venn diagrams showing the overlap in variant

calls between the graph-based and linear approaches for **(B)** SNVs and **(C)**INDELs. Values represent the number and proportion of variants unique to each method and shared between methods.

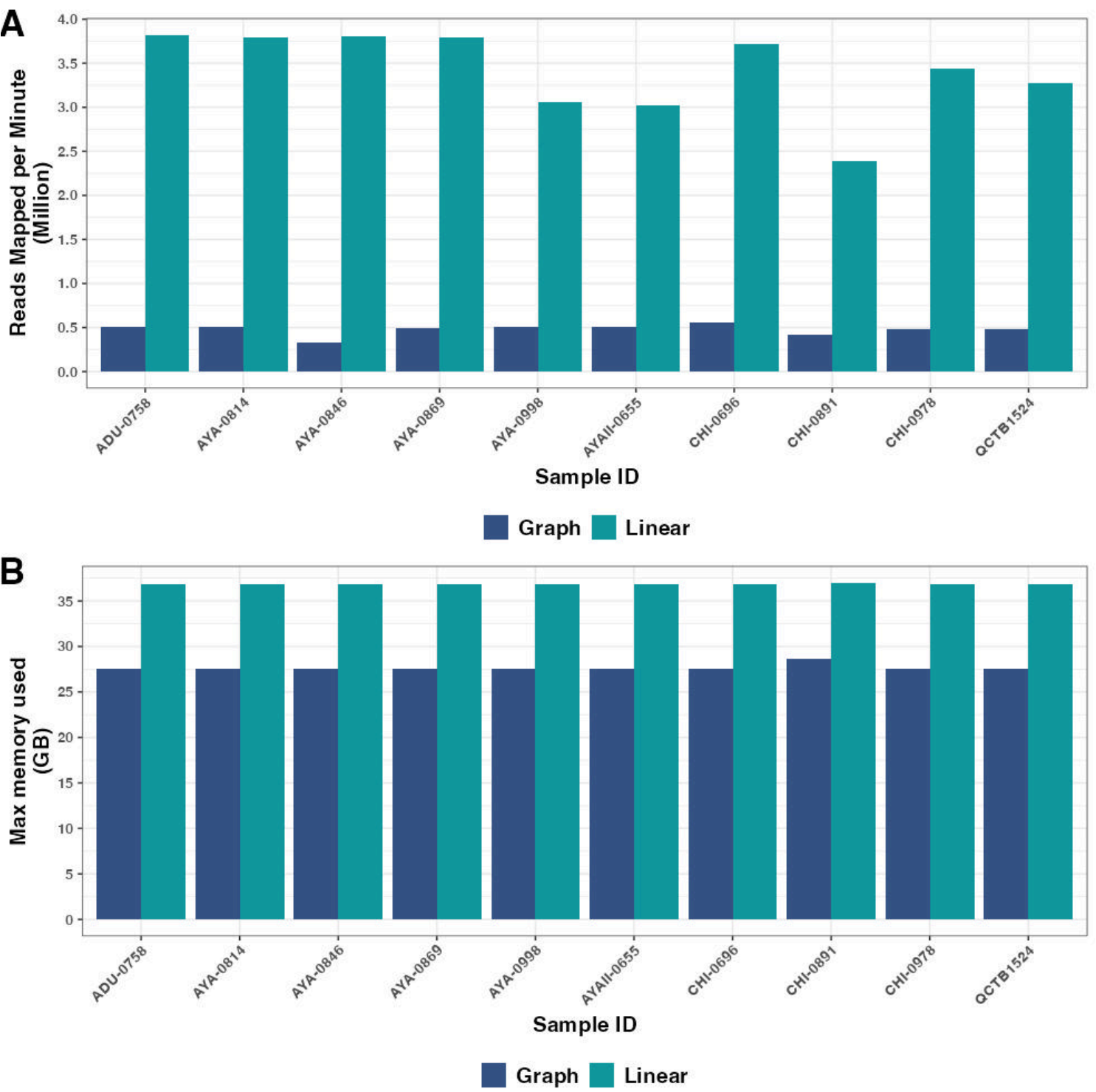


Figure 5

Computational performance comparison between graph-based and linear alignment approaches. **(A)** Reads mapped per minute (millions) for each of the 10 real patient samples, showing consistently higher mapping throughput for the linear approach compared with the graph-based method. **(B)** Maximum

memory usage recorded during analysis of each sample. Graph-based alignment required less memory across all samples relative to the linear reference workflow.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [AThompsonRobertsSuppData01.docx](#)