

Cancer genome standards for long-read sequencing using cancer cell line mixtures

Jia Zhang^{1,2}, Ho Yi Wong^{1,2}, Lingchen Liu^{1,2}, Lambros T. Koufariotis¹, Scott Wood¹,
Nadine Fitzpatrick¹, Jenny Quiatchon¹, Paul Collins¹, John V. Pearson¹, Nicola Waddell^{1,2,*}

¹QIMR Berghofer Medical Research Institute, Herston, Brisbane, QLD, 4006, Australia

²Faculty of Health, Medicine and Behavioural Sciences, The University of Queensland, Brisbane, QLD, 4102, Australia

*Correspondence address: Nicola Waddell. QIMR Berghofer Medical Research Institute, Herston, Brisbane, QLD, 4006, Australia. E-mail: nic.waddell@qimrb.edu.au

Abstract

Background: Long-read sequencing (LRS) improves genome alignment and facilitates resolving variants in genomic regions of low complexity, making it a promising approach for cancer variant detection and biomarker discovery.

Results: Here, we evaluate the performance of LRS using Oxford Nanopore sequencing to detect somatic variants across different tumour purities and sequencing depths by comparing to somatic variants from short-read sequencing. We generated experimental mixtures of cancer cell lines and matched normal cell lines to simulate 10 tumour purities (ranging from 0 to 100% tumour content). This resulted in 22 samples, which were sequenced using whole-genome LRS to a targeted depth of 60x, and somatic variants were identified. We down-sampled data to explore the optimal read depth for somatic variant detection. Our results show long-read variant calling tools achieve recall rates comparable to short-read gold standards. While tumour sequencing depth of 30x–60x is generally sufficient for detecting common variants, particularly structural variants, sequencing the matched normal sample at adequate depth is crucial for accuracy. Notably, we found variants unique to LRS that may represent real and previously undetected events.

Conclusion: This study highlights key factors for optimising cancer genome sequencing with LRS and provides a comprehensive dataset of cell line mixtures for future research and tool development.

Keywords long-read sequencing, Oxford Nanopore Technologies, somatic mutation, tumour purity, sequencing depth

Introduction

Cancer is a complex and multifaceted disease characterized by diverse genetic and epigenetic alterations, such as single-nucleotide variations (SNVs), structural variations (SVs), and changes in base methylation [1–3]. These genomic changes drive abnormal cellular proliferation and carcinogenesis [4], and represent targets for treatment [5] making the detection of mutations in tumour cells crucial for cancer research and precision therapy [6, 7]. While high-throughput short-read sequencing (SRS) has been widely used to identify somatic mutations across various cancer types, it often fails to detect variations in low-complexity, highly repetitive, or GC-biased regions due to limited mappability [8–10]. Long-read sequencing (LRS) technologies, such as those developed by Pacific Biosciences [11] and Oxford Nanopore Technologies (ONT) [12], offer extended read lengths that can span highly repetitive genomic regions effectively [13]. Initially used to enhance reference genome assembly and to resolve haplotypes [14, 15], LRS is increasingly applied in cancer genomics due to its ability to identify novel variations in previously unresolved regions of tumour genomes [16–20]. Recent studies using high-throughput platforms

like the Nanopore PromethION have enabled rapid and comprehensive profiling of various advanced cancer cohorts [20] and central nervous system tumours [21, 22], demonstrating significant potential for clinical applications. However, the use of LRS in cancer has largely been restricted to a small number of studies or the analysis of cell lines, which underscores the importance of benchmarking these technologies to optimize their clinical and research use.

The evolution of LRS technologies has led to a proliferation of approaches for analysing the resulting data [23], including base-calling, alignment, and variant calling. LRS requires computational methods that are distinct from SRS methods due to its longer read length and unique error profiles [24, 25]. Benchmarking studies have evaluated various LRS tools to analyse cancer genome data, including aligners, somatic structural variation detection, and methylation detection methods [26–28]. However, many of these evaluations relied on *in silico* data or sequence data from cell lines [29, 30], meaning they did not capture complexities of tumour tissue samples such as tumour purity. Notably, some benchmarking efforts are published alongside the development of new tools, which can introduce bias as optimal parameters may

Received: 6 May 2025. Revised: 16 October 2025. Accepted: 1 April 2026

© The Author(s) 2026. Published by Oxford University Press on behalf of GigaScience. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

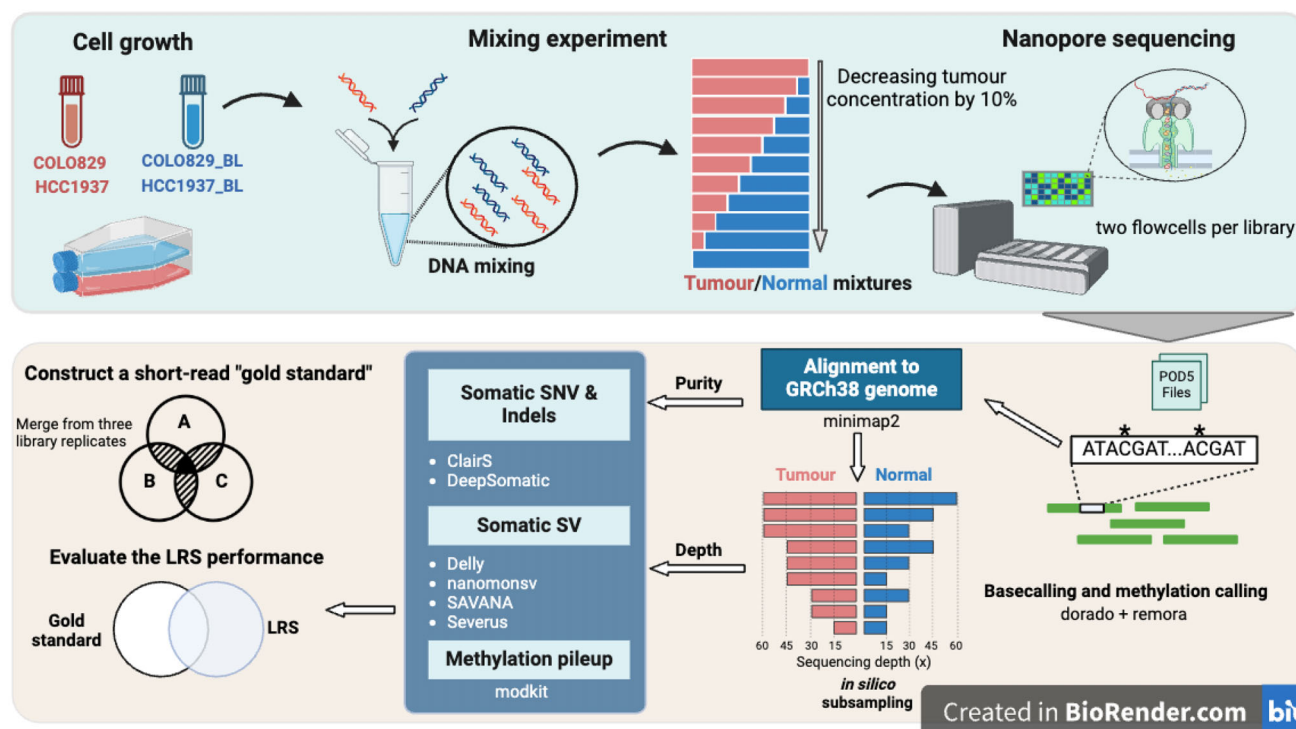


Figure 1 The overview of the experiment design. Two tumour-derived cancer cell lines, COLO829 and HCC1937, with their patient-matched non-tumour B lymphoblastoid (BL) cell line, were cultured and DNA extracted. The DNA was quantified, and tumour and non-tumour DNA samples were mixed to obtain tumour cell mixtures with decreasing tumour purity by 10% decrements. Libraries for each cell line mixture ($n = 11$ for each tumour cell line) were sequenced with the Nanopore PromethION sequencer with two flow cells per library. The raw data were basecalled with Dorado, and the 5mC methylation at CpG sites was called with a matched Remora model during basecalling. Sequence data were aligned to GRCh38 assembly and somatic SNVs, indels, and structural variations were identified for samples with different tumour purity levels. Sequencing depths were simulated by subsampling reads at 60x, 45x, 30x, and 15x for samples with 100, 80, 60 and 40% tumour purity levels. These combinations of sequencing depths ($n = 9$ per tumour cell line) were used for somatic SNV, indels, and SV detection. Short-read gold standard somatic mutations were constructed by overlapping high-confident sets from SRS of three replicate libraries for each cell line. Finally, the somatic variation results were compared with the short-read gold standard to evaluate the performance of LRS under different tumour purity and depths. The figure was created using BioRender (<https://www.biorender.com>).

not be used for each tool [31–34]. Therefore, independent benchmarking is valuable with the continuously emerging new tools and frequent version updates for existing tools [35].

Tumour purity and sequencing depth have a significant impact on the data quality and interpretation in tumour genomics studies [36–38]. Tumour purity refers to the proportion of cancer cells in a sample, and it can affect the signal-to-noise ratio, particularly for detecting somatic alterations (mutations specific to the cancer cells in a sample). Low tumour purity can mask somatic mutations by lowering the somatic variant allele frequency, potentially leading to false negatives [39]. Sequencing depth determines the sensitivity for identifying low-frequency variants as well as the ability to distinguish signals from germline cells; therefore, sequencing depth is particularly crucial in heterogeneous tumours [35] where low tumour purity and subclonal mutations are prevalent. Together, tumour purity and read depth influence the ability to detect somatic mutations, analyse clonal evolution, and understand tumour heterogeneity. Inadequate tumour purity or sequencing depth can result in the incorrect detection of clinically relevant alterations, which will negatively impact genomic discovery for cancer research and precision medicine [38, 40]. The implications of tumour purity and sequencing depth for LRS methods have not been explored.

In this study, we created cancer reference samples from two pairs of matched tumour-normal cell lines to measure the impact of tumour purity (0–100%) and sequence depth on somatic mutation detection. We sequenced the samples to an average read depth of $>60x$ using Oxford Nanopore LRS. We compared the somatic variant calls from different experimental settings against our SRS gold standard to determine optimal sequencing strategies. The results establish guidelines for expected data quality based on differing tumour purity levels and sequencing depths, providing valuable insights for optimizing experimental design in tumour sequencing studies. Furthermore, the generated dataset serves as a valuable resource for the development and validation of bioinformatic tools in this rapidly evolving field.

Results

Study design and overview

We generated two tumour-normal paired cancer samples to simulate different tumour purities and sequence read depths to evaluate the ability of LRS to identify somatic mutations (Fig. 1). We cultured two cell lines representing distinct tumour types

(melanoma, COLO829 and breast cancer, HCC1937) and a matched non-tumour cell line for each. DNA was extracted from each cell line and matched tumour- and non-tumour-derived DNA were serially mixed to simulate 10 tumour purity levels at 10% increments (0 to 100%). This resulted in 10 tumour mixtures and one matched blood lymphocyte (BL, as a germline/normal control) sample per cell line pair. A sequencing library was made for each sample and sequenced with the PromethION from ONT using two flow cells per library (R10.4.1). Basecalling was performed using Dorado (v0.5.1) with the super accuracy (SUP) model. The mean and median read quality scores were 21.8 and 23. The average yield per flow cell was 126 Gb, and the read length N50 was 10.4 kb (Supplementary Fig. S1A).

Two experiments were designed to assess the impact of common variables on somatic mutation calling in cancer samples: tumour purity and sequence read depth. To assess the impact of tumour purity, the 10 tumour cell mixtures and BL germline samples were sequenced to a mean depth of 70.61x for COLO829 (range 58.29x to 79.13x) and 61.57x for HCC1937 (range 53.42x to 71.14x) (Supplementary Fig. S1B), resulting in >95% of the genome being covered by at least 1 read (Supplementary Fig. S1C). The N50 (median aligned read length) for each library ranged from 9.4 to 10.4 kb (Supplementary Fig. S1D). We assessed impact of alignment using different human reference assemblies, CHM13-T2T and GRCh38. Slight improvements were observed using the CHM13-T2T reference with lower sequence divergence between reads and reference sequences and higher mapping rates (Supplementary Table S1). However, in our study, we used the GRCh38 assembly due to the comprehensive annotations and maximized compatibility with current tools and our short-read gold standard. To evaluate the performance of mutation detection under different sequencing read depths, we randomly subsampled sequence reads from four tumour mixtures (100, 80, 60, and 40%) per tumour type (COLO829 and HCC1937) and the matched BL sequence data to create tumour and germline BAM files with 60x, 45x, 30x, and 15x sequencing read depths. This resulted in nine combinations of sequencing depths.

In both experiments, we compared multiple approaches to call somatic single-nucleotide variants (SNVs), small insertions and deletions (indels), and structural variants (SVs). To assess the accuracy of the mutations detected with the LRS, we created a “gold standard” call from short-read data by sequencing DNA from the two tumour and matched non-tumour cell lines using three short-read replicate libraries. Sequencing resulted in a minimum read depth of 34x in BL and 67x in tumour samples, and we identified 40,422 SNVs, 416 indels, and 57 SVs in the “gold standard” short-read somatic calls for COLO829 and 48,982 SNVs, 3,236 indels, and 573 SVs for HCC1937 (Supplementary Fig. S2A–F).

Tumour purity impacts somatic mutation detection

Patient tumour tissues have different tumour purities; therefore, we created 10 tumour mixtures for each cell line to explore how tumour purity impacts somatic mutation detection from LRS. Two approaches were used for somatic SNV and indel detection from long-read data. ClairS [32] and DeepSomatic [31]. The shifted distributions of variant mutation allele frequency (VAF) for SNVs (Supplementary Fig. S3A) and methylation frequency at CpG sites that

are differentially methylated between the tumour and normal cell lines (Supplementary Fig. S3B) support the serial dilutions of tumour DNA by normal cell DNA.

Overall, the number of somatic mutations detected by ClairS and DeepSomatic decreased with decreasing tumour purity, resulting in a lower recall (Fig. 2A). Nevertheless, even at low tumour purity, the precision remained high (>79.5%). DeepSomatic had a higher precision than ClairS for COLO829 (ClairS range from 82.2 to 92.8%, DeepSomatic range 95.4 to 98.4%) and HCC1937 (ClairS range 79.5 to 86.1%, DeepSomatic range 84.5 to 93.5%) (Fig. 2A and Supplementary Fig. S4). In terms of recall, DeepSomatic had a higher mean recall at all tumour purity levels except the sample with 100% tumour purity compared to ClairS for COLO829 (ClairS 81.27%, DeepSomatic 86.39%); however, ClairS had a higher mean recall than DeepSomatic in HCC1937 in samples with lower tumour purity (<50%) (ClairS 81.43%, DeepSomatic 77.02%). To confirm these results, we performed a data simulation using the COLO829 data aligned to chromosome 22 with NanoSim [58] to introduce defined somatic SNVs to establish a ground-truth mutation set. This data confirmed that both ClairS and DeepSomatic show high (>99%) precision and recall at 100% tumour purity, with DeepSomatic achieving higher recall at lower purities than ClairS for this sample (Supplementary Table S3).

Within the sequenced mixtures (not simulated data), the majority of the SNVs were detected by both approaches, with 90% of SNVs being detected by ClairS and DeepSomatic for the COLO829 non-diluted sample and 87% for HCC1937 (Supplementary Table S2). The union of the two approaches achieved higher recall at the cost of introducing more FP events. The variant allele frequency of true positive (TP, present in “gold standard”) calls decreased with decreasing tumour purity for both COLO829 and HCC1937 (Fig. 2B) as expected. In contrast, the false positive (FP, absent from “gold standard”) calls for COLO829 do not show a decrease in AF, with FP having a greater range of AF in both COLO829 and HCC1937 at higher tumour purities (Fig. 2B). Interestingly, for ClairS, the read depth for FP calls was consistently lower than for TP calls (Supplementary Fig. S5A). This was not the case for DeepSomatic, suggesting that DeepSomatic FP variants might be TP variants missed by SRS meanwhile read depth could be used as a filter to remove real FP events in ClairS. The quality scores were lower for FP events compared to TP for both callers (Supplementary Fig. S5B). The different patterns observed in allele frequency, read depth, and quality score distributions between TP and FP events suggest that integrating these measures could serve as an efficient filter to reduce FP calls.

The average recall for indels (ClairS 71.62% for COLO829, 80.28% for HCC1937; DeepSomatic 57.85% for COLO829, 62.66% for HCC1937) (Fig. 2C) was lower than SNV (Fig. 2A). ClairS had a higher recall for indels at all tumour purity levels than DeepSomatic. Similar to SNV detection, the number of somatic indels detected by ClairS and DeepSomatic decreased with decreasing tumour purity, resulting in a lower recall (Fig. 2C). The average precision was significantly lower for indels (ClairS: 24.6% for COLO829, 42.39% for HCC1937; DeepSomatic: 68.55% for COLO829, 64.27% for HCC1937) than for SNVs (ClairS: 90.57% for COLO829, 83.61% for HCC1937; DeepSomatic: 96.87% for COLO829, 89.29% for HCC1937). The ClairS approach consistently had a lower precision in all tumour purities compared to DeepSomatic for both cell lines (Fig. 2C). There was no difference between the detection of insertion and deletion events (Fig. 2C).

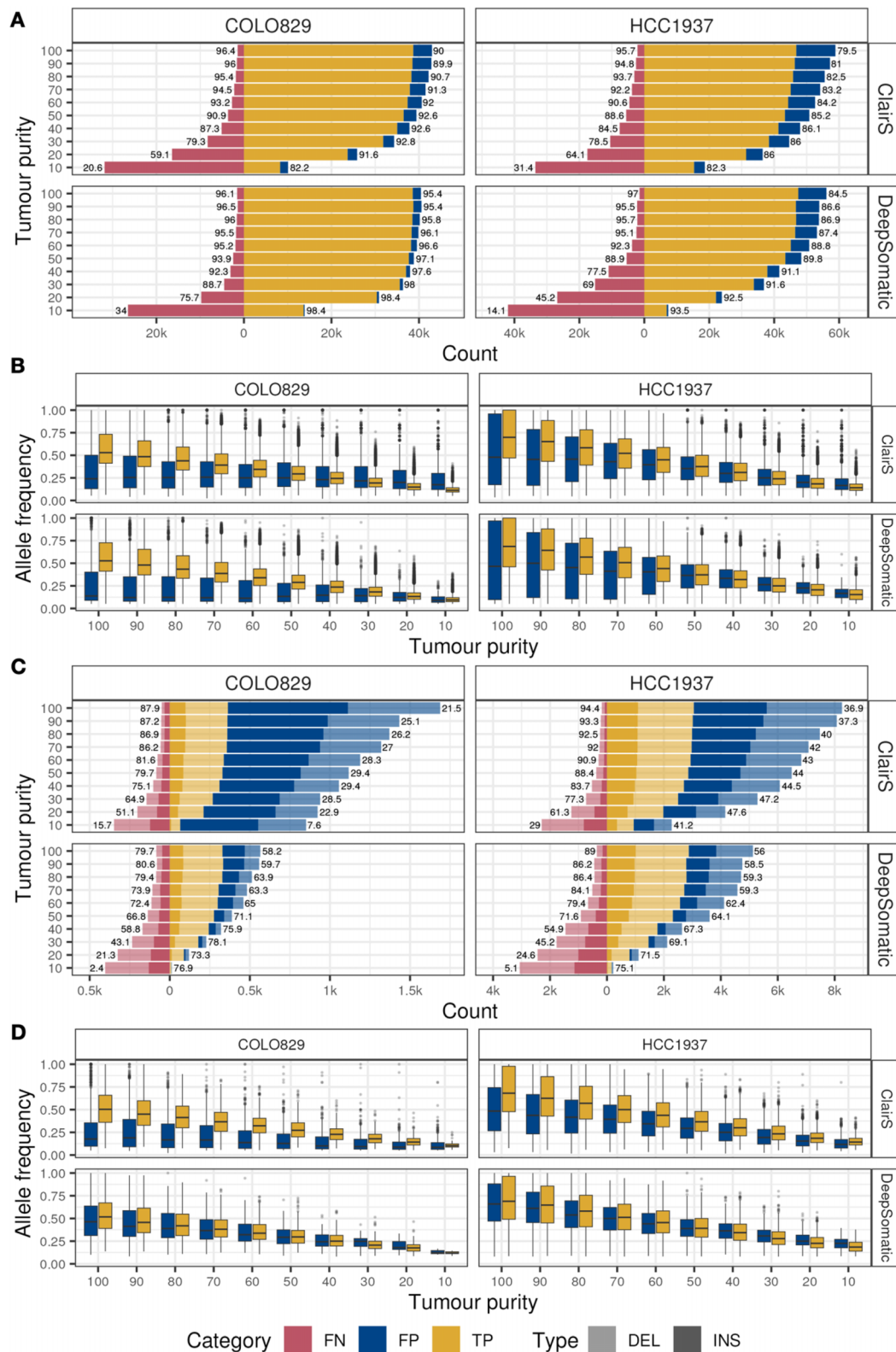


Figure 2 Tumour purity affects somatic SNV and INDEL calling. (A) Number of true positive (TP), false positive (FP), and false negative (FN) SNV calls in COLO829 and HCC1937 tumour cell mixtures using ClairS and DeepSomatic. The size of each call was present on the right side of 0 on the x-axis, and the missed call size on the left side. Precision and recall rates are labelled at the edges of the right and left sides of each row. (B) Boxplots comparing the distribution of variant allele frequency (VAF) for FP and TP in SNV calling with decreasing tumour purity. (C) The bar plot displays the number of TP, FP, and FN indel calls. The pale colour represents the deletions, and the solid colour shows the insertions. Precision and recall rates are labelled at the edges of the right and left sides of each row. (D) Boxplots comparing the distribution of VAF for FP and TP in indel calling with decreasing tumour purity.

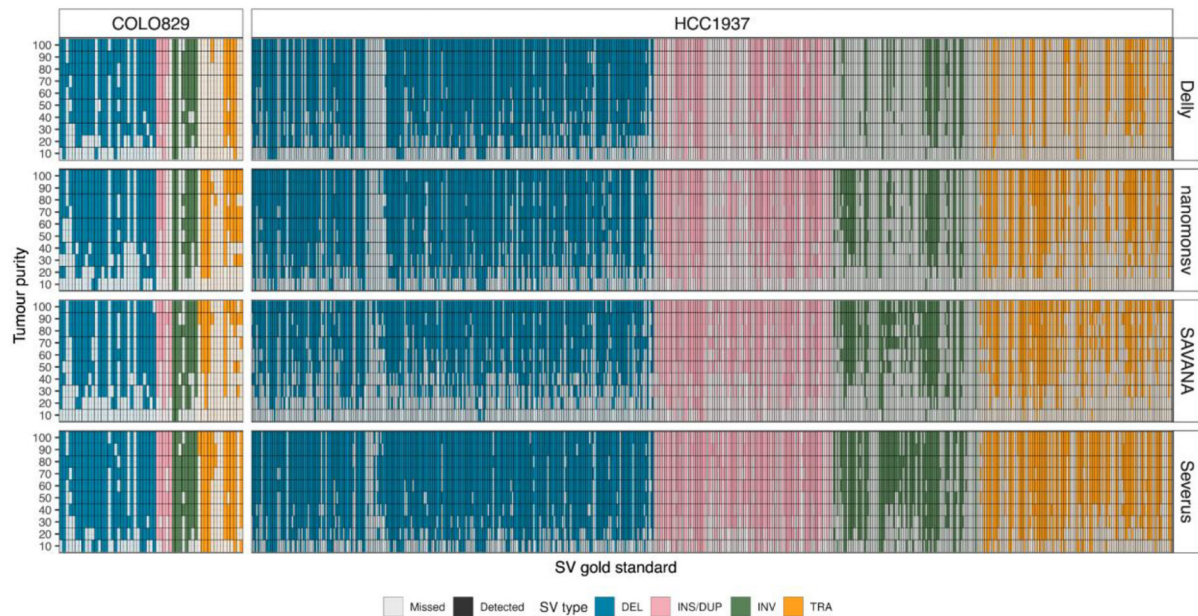


Figure 3 Tumour purity affects somatic SV calling. A heatmap of short-read “gold standard” SV events that are detected by long read using four SV callers in decreasing tumour purity levels for COLO829 and HCC1937 tumour cell lines. SVs of different SV types (DEL: deletion, INS/DUP: insertion/duplication, INV: inversion, TRA: translocation) are filled with different colours.

Similar to the SNVs calls, the mutation allele frequency of true positive (TP, present in “gold standard”) indels calls decreased with decreasing tumour purity for both COLO829 and HCC1937 (Fig. 2D). The AF of TP indels was higher than the FP calls for ClairS but not DeepSomatic, suggesting that AF could be used as a filter to remove FP indel events in ClairS. The read depth for FP and TP calls was similar for indel events in ClairS and DeepSomatic (Supplementary Fig. S6A), suggesting that read depth is not useful for determining TP events. The quality scores were lower for FP events compared to TP for both callers (Supplementary Fig. S6B), suggesting that quality could be used as a filter to remove potential FP calls.

We compared the SV calls under different purities using four SV detection tools in two cell lines. Most SV calls in the short-read “gold standard” are identified with long-read SV callers; however, with 3 out of 57 in COLO829 and 87 out of 573 HCC1937 short-read “gold standard” SVs were not detected by any LR tool in all tumour purity levels (Fig. 3), suggesting these events may be missed by LRS or false positive calls in the short-read data. Samples with higher purity usually generate more SV calls (Supplementary Fig. S7), with a decreasing SV count as purity decreases, particularly in HCC1937, where there are more SV events (573, compared to 57 in COLO829). The Severus tool displays the best recall rates (36.84–84.21% in COLO829 and 26.18–78.53% in HCC1937) at all tumour purity levels for both cell lines. Even at 20% tumour purity, the recall for Severus is 63.16% for COLO829 and 50.79% for HCC1937. In contrast, the recall for other tools at 20% tumour purity for COLO829 and HCC1937, respectively, was 42.11 and 36.47% for Delly, 40.35 and 35.25% for nanomonsv, and 28.07 and 20.59% for SAVANA. The recall rate for SAVANA was comparable to nanomonsv in higher purity samples (>60%); however, SAVANA did not perform as well in lower purity samples (Supplementary Fig. S7).

Sequencing depth affects somatic mutation calls

We sub-sampled tumour and germline samples to simulate different read depths (60x, 45x, 30x, and 15x) and then compared the somatic mutation detection under four purities (100, 80, 60, and 40%). This resulted in nine scenarios of tumour-germline sequencing depths under four tumour purities (Fig. 1).

Lower read depth in the germline sample is expected to impact somatic SNV detection. For ClairS, this resulted in an increase in recall and a decrease in precision, whereas for DeepSomatic, it resulted in a lower recall and a higher precision. When the germline samples were sequenced at 15x sequencing depth, the performance of these approaches was severely impacted, a finding that was confirmed in simulated data (Supplementary Table 3). Within the sequenced mixtures (not simulated data), using the samples with 100% purity, the precision dropped below 65% in both cell lines for ClairS, and the recall also dropped below 65% in HCC1937 for DeepSomatic (Fig. 4A). Furthermore, we observed the average allele frequency of FP SNVs when the normal sample at 15x depth higher than other read depth combinations (Supplementary Fig. S8). Similarly, in indel calling, lower sequencing depth in normal samples leads to decreased precision but increased recall for ClairS except when normal samples had only 15x read depth, whereas DeepSomatic results in reduced recall but improved precision (Fig. 4B). Interestingly, ClairS demonstrates robustness to the impact in the HCC1937 cell line, while DeepSomatic shows greater robustness in COLO829. Similar results were observed for samples with lower tumour purity (Supplementary Fig. S9).

Sequencing depth in both tumour and normal samples had a smaller impact on SV calling than SNV and indel calling,

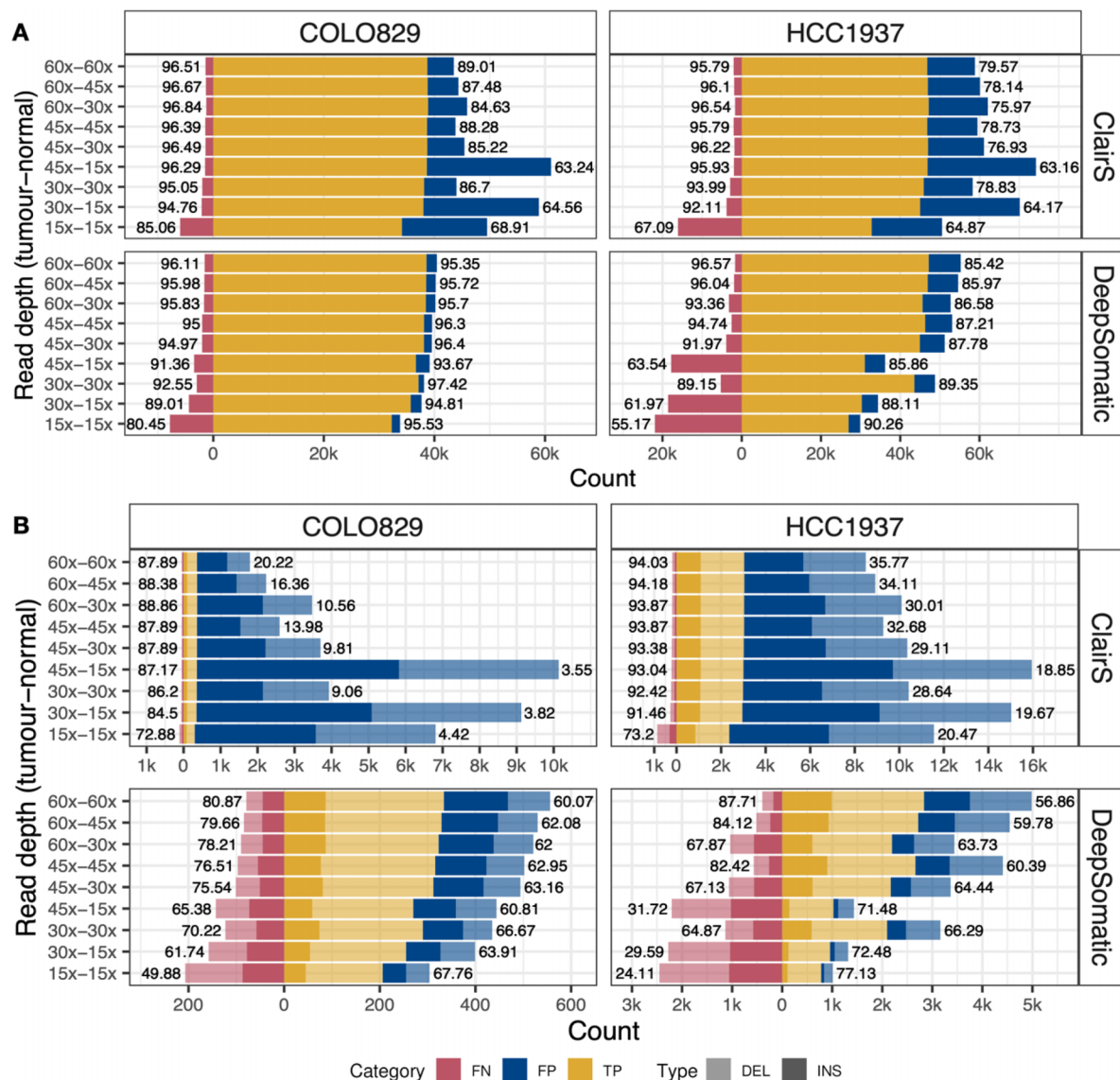


Figure 4 Sequencing depth affects somatic SNV and indel calling. Number of true positive, false positive, and false negative somatic variant calls under nine depth combinations (y-axis) of tumour and normal samples in COLO829 and HCC1937 using ClairS and DeepSomatic in samples with 100% tumour purity. (A) Somatic SNV calling performance in recall and precision of ClairS and DeepSomatic. (B) Somatic indel calling performance in recall and precision of ClairS and DeepSomatic, deletions and insertions were filled with stacked bars in different alpha.

particularly for nanomonsv and Severus (Supplementary Fig. S 10). Nanomonsv consistently demonstrated the highest precision in four tools across tumour purities, underscoring the importance of incorporating a reference panel to achieve accurate variant detection. Meanwhile, SAVANA is more sensitive not only to tumour purity but also to sequencing depth. In terms of recall, Delly outperformed SAVANA in recall when germline samples were sequenced at 15x, although its recall remained lower than that of nanomonsv and Severus. Severus exhibited the highest recall among the four tools across all depth combinations and detected more LR-specific (FP) SVs than nanomonsv, particularly in scenarios with lower depth in the normal sample. While the LR-specific SVs may represent novel SVs, they could also be read false positives.

False positives detected by LRS might be true

Despite the high recall rate in variant detection using LRS compared with SRS, we found that many LR-specific false positive calls might be real. The proportions of single substitution signatures (SBS) catalogues extracted in the false positives had high cosine similarity to our short-read gold standard with 0.79 and 0.89 for ClairS, 0.96 and 0.97 for DeepSomatic in COLO829 and HCC1937 (Fig. 5A). Notably, ClairS detected more T to C (A to G) type mutations with a context AA existing in FP-positive calls for both cell lines. We also found that false positive SNVs from ClairS tend to cluster along the centromere and telomere regions (Supplementary Figs S11 and S12) along chromosomes for both cell lines com-

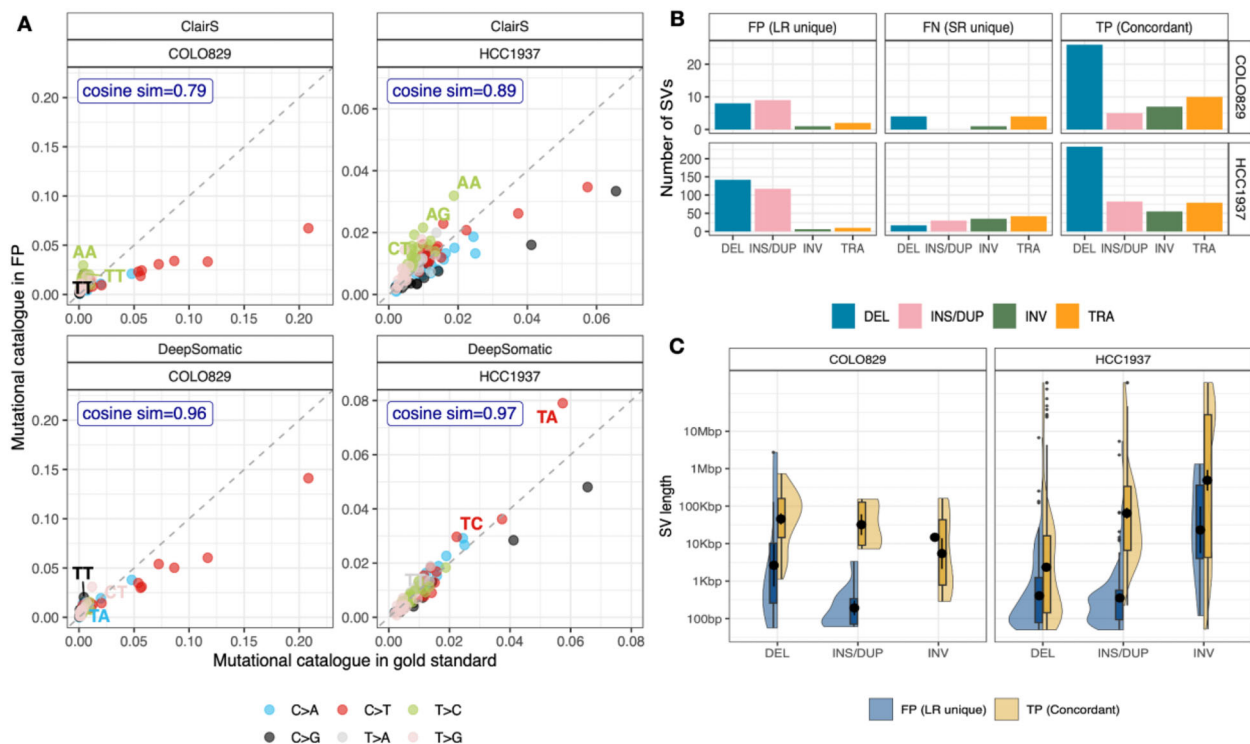


Figure 5 Potential novel events detected by LRS. (A) The comparison of the 96 mutational catalogues between the short-read gold standard and FP from LRS. (B) The number of SVs in four types (DEL: deletion, INS/DUP: insertion/duplication, INV: inversion, TRA: translocation) from FP, FN, and TP. (C) The SV length distribution of FP and TP events in DEL, INS/DUP, and INV types, the black dot in each bar indicates the mean value.

pared with DeepSomatic (Supplementary Figs S13 and S14). This suggests that DeepSomatic has more stringent variant calling at difficult genome regions. The genome stratification of false positive SNVs shows that ClairS calls more FPs in tandem repeats and low mappability regions, while DeepSomatic calls more FPs in areas of high or low GC%. Meanwhile, the genome stratification of false positive indels displays a similar pattern between the two tools with a higher proportion of FPs overlapped with tandem repeats and genomic regions with GC percentages <30% or > 55% (Supplementary Fig. S15).

The high-confidence LR SV calls presented in at least two out of four tools also contain many LR-specific SVs. While these SVs may appear as false positives, we expected some of these SVs to be real due to the technical advantage of long-read over short-read. In support of this, we found that a substantial fraction of the LR-specific SVs are located at genomic regions of extreme GC content (GC% <30 or >55) or within low mappability for short reads (Supplementary Fig. S16). Furthermore, these events are mostly deletions, insertions, and duplications with shorter lengths (<10 kb) (Fig. 5B, C), which means they can be sequenced by a single long read with high confidence. Meanwhile, SRS might miss these smaller SVs (<10 kb) due to the restriction of the length of reads and the uncertainty of the pair-end library sizes.

Germline leakage is more related to the sequencing depth but not tumour purity

Germline leakage refers to germline mutations being misclassified as somatic calls in the variant detection pipeline [41]. We estimate the germline leakage rate by quantifying somatic calls with a max-

imum population frequency above 0.1 in gnomAD (version 4.1) [42]. While this threshold may not capture all germline variants, we used these values to compare the germline leakage across tumour purity and sequencing depth. We found that the germline leakage rates remain low across tumour purity and start to increase with tumour purity less than 30% for ClairS while decreasing for DeepSomatic (Supplementary Fig. 17A). Notably, we found that the leakage spiked when the sequencing depth in the germline was 15x, suggesting enough depth in germline samples is crucial for avoiding germline variation contamination in somatic mutation detection (Supplementary Fig. 17B).

Benchmarking SV using publicly available gold standard

Using short-read data to build the SV gold standard is a challenge due to technical limitations of SRS to detect the full set of SVs in the genome. Previously, a COLO829 SV truth set SV, including 68 SVs, was thoroughly characterized with multiple sequencing platforms, including short-read and long-read technologies [43]. We benchmarked our long-read SVs to this truth set to evaluate how many SVs that were missed by our SR gold standard can be detected by LRS. The recall rates are slightly lower than the results when benchmarking with our SR gold standard (Supplementary Fig. S18). This suggests our “gold standard” captured most SVs; however, differences exist because our “gold standard” was constructed using single-platform sequencing data and also genetic changes may accumulate in cell line passing causing cells from different organizations to increasingly differ across time.

Discussion

In this study, we sequenced a series of *in vitro* cell line mixtures with different tumour purity levels from two cancer types and compared the performance of long-read somatic variant detection tools, including SNVs, indels, and SVs (≥ 50 bp). Additionally, our *in silico* sequencing depth down-sampling allowed us to evaluate the impact of sequencing read depths of tumour and normal samples on variant detection. Our results provide a practical guide for cancer genomic sequencing of clinical samples, and the dataset of sequenced mixtures could be valuable for benchmarking the robustness of variant calling tools under different tumour purity settings.

Previous studies have benchmarked LRS for variant calling in a range of genomic contexts, but many focused on germline-specific variants or prioritized specific genomic features such as SVs [27], CpG methylation [44], and mRNA [45], some with the impact of tumour purity addressed by the *in silico* mixing dataset [29]. However, there remains a critical need to benchmark LRS for variant calling in the context of tumour genomics. Unlike germline variants, somatic mutations in tumours exhibit substantial heterogeneity, both within and across cancer types, posing unique challenges for accurate detection [46]. Through experimental mixing DNA of tumour and normal cell lines and *in silico* down-sampling data, we showed that LRS tools can detect $>50\%$ of the variants of all kinds from a short-read “gold standard” at tumour purities as low as 20% and for 30x–60x sequencing depth in tumour. Importantly, our data show that having enough depth in germline sample ($>15\times$) is critical for recall and precision of somatic mutations. These findings can guide parameters for clinical sequencing protocols.

As expected, tumour purity significantly affects somatic mutation detection, with lower purity reducing sensitivity across SNVs, indels, and SVs. This highlights the challenge of identifying rare somatic alleles in samples with significant stromal cell content. However, even at low tumour purity (e.g., 20%), LRS, especially with Severus for SVs, maintains a reasonable recall rate, demonstrating its utility in low-cellularity samples. Although Nanopore sequencing has lower base qualities (Q20–Q30) than the current SRS (Q30–Q40), deep-learning-based SNV callers like ClairS and DeepSomatic effectively mitigate sequencing errors. ClairS combines pileup and full-alignment models using Bi-GRU and ResNet architectures to classify variants as germline, somatic, or artefacts [32]. DeepSomatic, using a CNN-based model, processes alignment pileups as images [47]. These architectural differences likely drive performance differences. DeepSomatic’s higher precision for SNVs and indels suggests a more conservative, error-resistant approach, possibly due to its training data and algorithm design. While ClairS achieves comparable SNV recall and better indel recall, it has lower precision, suggesting it requires stringent downstream filtering of variants based on biological context. Notably, ClairS detects variants in centromeric and telomeric regions, indicating its potential to identify mutations in genomic regions traditionally inaccessible to short reads. These results suggest that despite lower base qualities, LRS is able to accurately call somatic mutations even in samples with low tumour content.

Among the four long-read SV detection tools evaluated, Severus consistently outperformed Delly, nanomonsv, and SAVANA across varying tumour purities and sequencing depths. Severus maintained high recall even at low tumour purity and sequencing depth

($15\times$), suggesting its ability to detect tumour-specific signals with minimal interference from germline contamination. SAVANA’s performance declined at lower tumour purities and showed high sensitivity to reduced sequencing depth, potentially limiting its applicability in clinical settings. However, its recent integration of tumour purity, ploidy, and copy number alteration inferences may address these problems [34]. Nanomonsv performed similarly to SAVANA but remained more robust at lower tumour purities and sequencing depths. Its use of an internal human germline SV reference panel makes it more conservative in SV detection, which aligns with previous findings [27, 48].

LRS inherently offers advantages for SV detection over short-read approaches. Given that SNV and small indel detection show comparable performance [49] between long- and short-read technologies [50], a key question is whether LRS enhances SVs discovery by detecting variants missed by SRS. While previous studies suggested that most large SVs in tumour genomes are detectable by short reads, our findings indicate that long reads uniquely capture numerous short-to-intermediate SVs (<10 kb), which were underrepresented in earlier analyses [51]. This further underscores the importance of LRS in tumour genomics, particularly for identifying novel SV markers beyond the resolution of short-read methods.

To enable future cancer genomics studies and translate LRS into the clinic, we need to know what read depth is required to reliably detect somatic mutations. Our sequencing depth down-sampling experiment suggests that balancing coverage between tumour and normal samples can improve somatic variant detection potentially by minimizing germline contamination. Furthermore, sequencing depths of 15x or less for the germline sample significantly harm somatic variant calling and should be avoided. In our study, we sequenced both tumour and normal samples using two flow cells aim for a read depth of 60x. While we obtain data of $\sim 30\times$ genomic sequencing coverage per flow cell, a study has shown that a single Nanopore PromethION flow cell can generate $>33\times$ human genome coverage with ultra-long reads (N50 >35 kb) with optimized DNA processing and library preparation [52]. Assuming that sequencing with a shorter (10 kb) read length N50 produces higher yields per flow cell [49], in the future we can reduce sequencing costs by using only one flowcell each for tumour and normal.

Although long reads allow improved mapping in genomic regions of low complexity, the alignment accuracy remains imperfect [53]. Some suggest the reason to be low read accuracy; however, the main factor could be the genome assembly [54]. We used the GRCh38 assembly in this study as it provides more comprehensive clinical annotations for the downstream analyses; however, GRCh38 remains incomplete with unresolved gaps and repetitive sequences of unknown size. Recent studies suggest that using the CHM13-T2T [14] assembly results in better genetic variant detection from more precision alignment, especially in highly repetitive regions such as centromeres and telomeres [55, 56]. Future studies exploring the impact of reference assemblies on variant calling performance may provide valuable insights into cancer genome sequencing.

The “gold standard” used in this study was constructed with the SRS with all the limitations that imply. Previously, researchers have utilized the synthetic mixtures of two fully homozygous human cell lines to mock somatic mutations [57], and apply read simulators such as NanoSim [58] or pbsim2 [59] to upsample data for

validation of variant detection. While this is a good strategy to explore variant calling performance under extreme scenarios, differences in variation calling performance were found when using the real tumour cell line [33], which underscores the heterogeneous nature of cancer samples [60]. A major output of our study is the creation of reference data (with a range of tumour purities) that can be used by research community. To this end, we deliberately selected two cancer types, one with a high SNV burden, the other with a high SV burden. Establishing references for other key features of cancer genomes, such as subclone mutations, tandem repeats, and complex SVs, will largely facilitate the development of better computational tools [61, 62].

This study explored the optimal experimental design for cancer genome sequencing using LRS through creation of tumour purity mixtures. Many LRS analysis tools are still under active development, and while we are far from a definitive “best practice” for cancer variant calling using LRS, we have investigated key aspects key and highlighted directions for improvement. Importantly, we provide a publicly available dataset of LRS cancer cell line mixtures, offering a valuable resource for future tool development and benchmarking in this fast-evolving field.

Materials and methods

Ethical statement

This work involves the use of human cell lines, which were approved by the QIMR Berghofer Human Research Ethical Committee (P3577 and P3527).

Cell culture, DNA extraction, and mixing

COLO829 (RRID:CVCL_1137), COLO829_BL (RRID:CVCL_1999), HCC1937 (RRID:CVCL_0290), and HCC1937_BL (RRID:CVCL_3281) were maintained in RPMI +10% FBS. All cell lines were tested for mycoplasma and STR profiling. Cells were harvested between passages 8 and 10. Cells were dissociated using 0.25% trypsin at 37°C for 5 min and spun at 1,200 rpm for 5 min. Cells were washed once with PBS and kept at −80°C until DNA extraction. DNA was extracted using AllPrep DNA/RNA Mini Kit (802024, Qiagen) according to the manufacturer’s protocols and eluted in TE buffer. To prepare samples with various tumour purities, both tumour and normal DNA were diluted to 50 ng/μl using the TE buffer. The tumour and its corresponding normal DNA were mixed in specific ratios to create 10 samples that represent tumour purities ranging from 0 to 100% with increments of 10%.

Library preparation and Oxford nanopore sequencing with PromethION

DNA quantity and quality were assessed using an Implen NanoPhotometer, then run on a TapeStation (Agilent) to evaluate fragment size. High molecular weight DNA was sheared to 10–12 kb using Covaris g-tubes. Approximately 1 μg of sheared genomic DNA was used to make libraries using the ONT ligation sequencing kit (SQK-LSK114), according to the manufacturer’s instructions, except that ligation was extended to 30 min. Each library was loaded at approximately 10 fmol on two separate R10.4.1 PromethION flow cells and sequenced for 72 h on a

PromethION P24 sequencer within QIMR Berghofer Scientific Services.

Basecalling, methylation calling, and alignment

The raw POD5 files from the sequencer were basecalled using Dorado (v0.5.1) [63] with the super accuracy (SUP) model for the R10 flow cell (dna_r10.4.1_e8.2_400bps_sup@v4.3.0). The 5mC and 5hmC methylation were also detected using Dorado with the matched Remora model (dna_r10.4.1_e8.2_400bps_sup@v4.3.0_5mCG_5hmCG@v1). SAMtools (v1.17) [64] was used to convert the unaligned basecalled BAM file from basecalling to FASTQ format and filter out reads with average quality (QS) <10 before alignment, and minimap2 (v2.26) [65] was used to align reads to the GRCh38 genome assembly (−y −MD −ax map-ont). The aligned BAM files were sorted and indexed using SAMtools. The percentage of 5mC methylated reads at CpG sites was calculated using modkit (v2.4.0) pileup based on the MM and ML tags of reads from both strands in the BAM files (−combine-strands −cpg).

Read depth simulation

To evaluate the performance of variant calling under sequencing strategies, we used Sambamba (v1.0.1) [66] to downsample data from the tumour and the matched B lymphoblast BAM files with 60x, 45x, 30x, and 15x read depths in samples with 100, 80, 60, and 40% tumour purities. Reads were randomly sampled from the aligned BAM files at the proportion based on the expected and original depth. If a BAM file did not reach a 60x read depth, we used the original BAM file (53.42x for the library of HCC1937 cell line at 40% tumour purity). In total, we generated nine combinations of sequencing depths with either higher read depths in tumour samples (60x-45x, 60x-30x, 45x-30x, 45x-15x, 30x-15x) or equal read depths in both tumour and normal (60x-60x, 45x-45x, 30x-30x, 15x-15x). The depth of subsampled BAM files was confirmed with mosdepth (v0.2.9) [67] to check the potential effect of varied read lengths.

Somatic SNV, indel, and SV detection

We used ClairS (v0.1.7) and DeepSomatic (v1.6.0) for SNV and short indel detection in tumour cells. These tools were selected as they are both actively updated and designed to call somatic SNVs and indels using a pair of tumour and normal BAM files to predict tumour-specific variants using pre-trained deep-learning models. For ClairS, we used the ont_r10_dorado_sup_5khz model for somatic SNV detection and the r1041_e82_400bps_sup_v430 model from rerio for Clair3. For DeepSomatic SNV calling, we used default parameters with an internal ONT model provided by Google Health. The final results in VCF were filtered for somatic SNVs with GT=”1/1” and FILTER=”PASS” using BCFtools (v1.19) [68] as described in the DeepSomatic document.

We used four tools (Delly (v1.2.6) [69], nanomonsv (v0.7.1) [70], SAVANA (v1.0.5) [34], and Severus (v1.0) [33]) to call SV using long reads. Each tool uses a pair of tumour and normal BAM files and runs with default parameters with SV length less than 50 bp were excluded. In addition, nanomonsv included an extra refer-

ence panel of 30 human genomes to filter somatic SVs. As part of the nanomonsv pipeline, SV events that overlapped with simple repeats were filtered and the SV type was annotated using the `sv_type.py` script of the package. Phased and haplotagged BAM files were provided to Severus, and variable number tandem repeat regions were used in the analysis. WhatsHap (v1.4) [71] was used to phase genomics variants in germline VCF files and add tags to each read in BAM files with known haplotypes.

Short-read sequencing for cancer cell lines

DNA was extracted for COLO829 and HCC1937 cells and matched B lymphoblastoid cells using the Qiagen AllPrep DNA mini kit according to the manufacturer's protocol (Qiagen, Germany). Whole genome analysis was performed on DNA from tumour and BL cells using DNA PCR-free prep and sequenced on a NovaSeq (Illumina).

The short-read gold standard for SNVs and indels

We used the concordant high-confident somatic SNV calls based on three high-depth short-reads sequencing of COLO829 and HCC1937 as our “gold standard.” The cell lines were cultured from the same purchase within a few passages and thus can avoid extensive sub-clonal somatic mutations. The short-read data of three biological replicates were analysed with our in-house standard Cromwell pipeline to detect somatic SNVs with GATK (v4.0.4.0) [72] and qSNP (v2.1.4) [73] and somatic small indels with GATK. For merging SNVs, we first split multi-base variants into consecutive SNVs and also split multiallelic sites into multiple entries using the BCFtools norm module. Next, the intersection of high-confidence SNVs for three libraries was calculated using the BCFtools isec command. Finally, the three VCF files were merged and annotated with the intersection results using BCFtools merge. SNVs that are detected in at least two libraries were used as a short-read “gold standard.” The indel gold standard was generated similarly to SNV but without atomization.

The short-read gold standard for SVs

For each library of COLO829 and HCC1937, we used Delly, GRIDSS (v2.13.2) [74], and LUMPY (v0.3.1) [75] to call SV from short-read data. The somatic SVs were extracted from each tool and merged within each library. The SVs that were detected by at least two tools were further merged between libraries for each cell line. Finally, SVs that were shared by at least two libraries were used as the SV gold standard. The merging of SVs is not as straightforward as SNVs. The output from each tool was first converted into a simple VCF format with five SV types (INS, INV, DEL, DUP, and TRA), and Jasmine [76] was used to merge multiple VCF files and to determine the overlapping.

EPIC array data analysis

The intensity data (IDAT) files of each cell line were loaded and analysed with the minfi (v1.48.0) [77] R package. Specifically, the raw signals were transformed and normalized as beta values at each probe CpG position, and the probes that contained an SNP at the CpG interrogations or the single-nucleotide extension were

excluded. To compare long-read methylation with EPIC array data, we performed liftover for EPIC array results from hg19 to hg38 using R packages GenomicRanges (v1.54.1) [78] and rtracklayer (v1.62.0) [79] with the chain file from the UCSC Genome Browser [80]. The methylation data were also used to confirm the tumour purity levels within the serial diluted samples. Specifically, differentially methylated CpG sites between the tumour and blood lymphocyte cell lines were determined as probes with beta values ≥ 0.7 in tumour and ≤ 0.3 in BL, or ≤ 0.3 in tumour and ≥ 0.7 in BL. The distribution of these CpG sites was then assessed in the samples with different tumour purities for COLO829 and HCC1937.

Benchmarking

We benchmarked the somatic SNVs and indels calling using the Python script `compare.py` of ClairS. We also used `hap.py` for comparison, which generated similar results. All SNVs and indels with PASS filter were used as the target to calculate the recall, precision, and F1 values with the corresponding gold standard for each cell line and variant type as the truth set. For SV benchmarking, tools had different formats to represent SV events. We converted the breakpoint pairs from different tools into the simple type VCF format and merged them with the short-read gold standard using Jasmine with default parameters, and only SVs with a length greater than 50 bp were considered. The public COLO829 SV truth set (Version v4) was downloaded from Zenodo (10.5281/zenodo.7515830) and compared with the results for a 100% tumour purity sample. The recall and the precision rates were calculated as

$$\text{Recall} = \frac{\text{overlap calls between LRS and SRS}}{\text{total calls from SRS}},$$

$$\text{Precision} = \frac{\text{overlap calls between LRS and SRS}}{\text{total calls from LRS}},$$

$$\text{FN} = \text{unique calls in SRS},$$

$$\text{FP} = \text{unique calls in LRS}.$$

Genome regions and stratifications

To investigate the genomic distribution of variants identified as TP, FP, and FN, genome stratifications for GRCh38, including low complexity, segmental duplications, low mappability, and GC content outliers, were downloaded from the genome-in-a-bottle consortium (v3.4) [8]. The overlapping status of each variant in regions of different categories was determined in R with the GenomicRanges and VariantAnnotation (v1.48) [81] packages. The genome-wide distribution of TP, FP, and FN SNVs was visualized using rainfall plots, which were aligned with chromosome ideograms annotated with cytogenetic banding patterns (gieStain) obtained from the UCSC Genome Browser [82].

Mutational signatures analysis

All SNV somatic mutations were converted into a 96-channel mutational profile (SBS-96 vector) and were subsequently decomposed into contributions of COSMIC reference catalogue (v2.0, GRCh38) using the R package `deconstructSigs` (v1.8.0) [83]. To evaluate whether LR-specific variants, classified as false positives, might represent genuine mutational events, we compared their

mutational spectra with the short-read gold standard set. Specifically, a cosine similarity was computed between the two profiles.

Germline leakage estimate

To assess the extent of germline contamination in somatic variant calling, we quantified the proportion of SNVs commonly observed in the general population. Population allele frequencies were retrieved using BCFtools, which was used to query each variant against the gnomAD VCF files (version 4.1) [42]. SNVs with a population allele frequency (AF_grpmax) greater than 0.1 were classified as germline variants.

Visualization and IGV check

Data wrangling was implemented in R (v4.3.1) with tidyverse (v1.2.1) [84] and visualized using ggplot2 (v3.4.4) [85] and ggh4x (v0.2.8) [86]. The long-read- and short-read-specific SV events were extracted as BEDPE files, and the IGV view of regions was generated using igv-reports (v1.12.0).

An *in silico* simulation with simulated COLO829 Nanopore sequencing data

Chromosome 22 was extracted from GRCh38 genome and integrated with gold standard germline SNVs and somatic SNVs of COLO829 and COLO829_BL using SAMtools and BCFtools. Nanopore reads were simulated using NanoSim (v3.2.3) trained on a subset of empirical data of COLO829_BL to approximate realistic read length and error characteristics. Matched tumour-normal datasets were generated based on the synthetic tumour and germline genomes at tumour purity 10, 20, 40, 60, 80, and 100% and sequencing depths of 15x, 30x, 45x, and 60x, respectively. These datasets were processed using the same alignment and variant-calling workflow as the empirical mixtures, and performance was evaluated against the known truth set.

Availability of source code and requirements

Project name: nanopore_celllines_benchmark.

Project homepage: https://github.com/bakeronit/nanopore_celllines_benchmark.

Operating system: Linux (Rocky Linux 9.7).

Programming language: Python, R, Bash.

Other requirements: Snakemake $\geq$ 8.2.0, tidyverse $\geq$ 2.0.0.

License: BSD-3-Clause license.

Additional files

Supplementary Figure S1. The quality of sequence data and alignment for tumour mixtures and matched normal samples. (A) Dot plot showing the total data yield (gb, gigabases) and read length N50 kb (kilobases) for each flow cell. Samples are ordered by sample type, then gb of sequence. (B) Total sequencing depth for each sample, with the x-axis showing the sample tumour purity and the y-axis indicating the sequence read depth. (C) The proportion of genome assembly covered by at least one read in each

sample. (D) Alignment length N50 for each sample. Samples are coloured by type (blue: B lymphoblast non-tumour and red: tumour mixture).

Supplementary Figure S2. Creation of short-read “gold standard” somatic mutation calls for the COLO829 and HCC1937 cell lines. The COLO829 and HCC1937 tumour-derived cell lines and their matched non-tumour cell lines were sequenced in triplicate in short reads. For COLO829, the UpSet plots show the number of somatic (A) single-nucleotide variants (SNVs), (B) insertion and deletion (indels), and (C) structural variant (SV) events detected uniquely in each replicate and shared by three or two replicates. For HCC1937, UpSet plots show the number of somatic (D) SNVs, (E) indels, and (F) SV events detected uniquely in each replicate and shared by three or two biological replicates. Blue bars represent events shared by at least two replicates, which were used to define the gold standard.

Supplementary Figure S3. The distribution of variant allele frequency (AF) and methylation in samples with different tumour purities. (A) The distribution of variant allele frequency of SNVs at the genomic regions with a copy number equal to two in COLO829 and HCC1937. The lines are coloured by tumour purity from 100 to 10%. (B) The methylation frequency distribution in CpG sites (x-axis) in tumour samples with different tumour purities (y-axis) for COLO829 (left) and HCC1937 (right). Only the CpG sites with differential methylation frequency between the 100% tumour and blood lymphocyte cell lines are included in the plot. Plots are coloured by the methylation status in the 100% tumour sample with a high (red) or low (blue) methylation level.

Supplementary Figure S4. Precision-recall curves (AUPRC) for variant detection across different tumour purities. The precision (y-axis) and recall (x-axis) of long read calls against the short read “gold standard” are shown for SNVs (A), indels (B), and SVs (B). Points are coloured by tumour purity. In panels (A) and (B), points in circles shape represent ClairS results, and triangles represent DeepSomatic. In panel (C), line colour indicates SV callers. The results for COLO829 (left panel) and HCC1937 (right panel) cell lines are shown.

Supplementary Figure S5. Read depth and quality scores for somatic SNV calls. (A) Box plot showing read depth in tumour samples with varying tumour purity for TP (yellow) and FP (blue) SNVs called by ClairS and DeepSomatic for COLO829 and HCC1937 cell lines. (B) Box plot of QUAL scores from ClairS and DeepSomatic for TP and FP SNV calls in samples with varying tumour purity for COLO829 and HCC1937. The line in the box is the median of the data, the whiskers extend to the minimum and maximum values within 1.5 times the interquartile range (IQR) from the quartiles. Data points beyond the whiskers are shown as individual outliers. FP: false positive; TP: true positive; QUAL: quality.

Supplementary Figure S6. Boxplots of read depth and quality scores for TP and FP indel calls in tumour samples. (A) Read depth in tumour samples with varying tumour purity for TP (yellow) and FP (blue) indels called by ClairS and DeepSomatic for COLO829 and HCC1937 cell lines. (B) QUAL scores from ClairS and DeepSomatic for TP and FP indel calls in samples with varying tumour purity for COLO829 and HCC1937. The line in the box is the median of the data, the whiskers extend to the minimum and maximum values within 1.5 times the interquartile range (IQR) from the quartiles.

Supplementary Figure S7. Precision and recall of SV calling using four tools in samples with varied tumour purity for COLO829 and

HCC1937. The bar height represents the total number of SV events detected by each tool. Bar colour indicates SV call concordance: blue for FP events; yellow for TP events; red for FN events. Precision and recall rates are labelled to the right and left of each bar, respectively.

Supplementary Figure S8. Variant allele frequency (VAF), read depth in tumour (DP), and quality score (QUAL) of FP and TP SNV calls from ClairS and DeepSomatic in COLO829 and HCC1937 samples with nine sequencing depth combinations. Panels are arranged with COLO829 on the left and HCC1937 on the right, and ClairS on the top and DeepSomatic on the bottom. Within each panel, results are further separated by sequencing depth combinations ranging from 15x–15x to 60x–60x (tumour–normal). FP calls are shown in blue and TP calls in yellow.

Supplementary Figure S9. Recall and precision of Indel calling across sequencing depths and tumour purities. Each panel displays the number of indels classified as true positives (TP, yellow), false positives (FP, blue), and false negatives (FN, red) across nine combinations of tumour-normal sequencing depths and four tumour purities (100, 80, 60, and 40%). Precision and recall are displayed to the right and left of each bar, respectively.

Supplementary Figure S10. Recall and precision of SV calling using four tools for nine different tumour-normal read depth combinations in COLO829 and HCC1937. Each of the eight blocks represents results from an SV caller for a cell line. Within each block, the y-axis indicates the read depth combinations, and the x-axis represents the number of discordant or concordant SV calls.

Supplementary Figure S11. The rainfall plots of FN, FP, and TP SNVs of ClairS in COLO829. Chromosomes 1–22 were listed as rows, with the ideogram shown at the top. The colour in the chromosome ideograms is shown in the key. The x-axis presents the position on each chromosome, and the y-axis presents the distance between an SNV and its neighbouring SNV. Points are coloured as FN (red), FP (blue), and TP (yellow). Cytogenetic bands (Stain): gneg (light), gpos25–100 (increasingly dark), acen (centromere, red), gvar (heterochromatic regions), and stalk (secondary constriction).

Supplementary Figure S12. The rainfall plots of FN, FP, and TP SNVs of ClairS in HCC1937. Chromosomes 1–22 were listed as rows, with the ideogram shown at the top. The colour in the chromosome ideograms is shown in the key. The x-axis presents the position on each chromosome, and the y-axis presents the distance between an SNV and its neighbouring SNV. Points are coloured as FN (red), FP (blue), and TP (yellow). Cytogenetic bands (Stain): gneg (light), gpos25–100 (increasingly dark), acen (centromere, red), gvar (heterochromatic regions), and stalk (secondary constriction).

Supplementary Figure S13. The rainfall plots of FN, FP, and TP SNVs of DeepSomatic in COLO829. Chromosomes 1–22 were listed as rows, with the ideogram shown at the top. The colour in the chromosome ideograms is shown in the key. The x-axis presents the position on each chromosome, and the y-axis presents the distance between an SNV and its neighbouring SNV. Points are coloured as FN (red), FP (blue), and TP (yellow). Cytogenetic bands (Stain): gneg (light), gpos25–100 (increasingly dark), acen (centromere, red), gvar (heterochromatic regions), and stalk (secondary constriction).

Supplementary Figure S14. The rainfall plots of FN, FP, and TP SNVs of DeepSomatic in HCC1937. Chromosomes 1–22 were listed as rows, with the ideogram shown at the top. The colour in the

chromosome ideograms is shown in the key. The x-axis presents the position on each chromosome, and the y-axis presents the distance between an SNV and its neighbouring SNV. Points are coloured as FN (red), FP (blue), and TP (yellow). Cytogenetic bands (Stain): gneg (light), gpos25–100 (increasingly dark), acen (centromere, red), gvar (heterochromatic regions), and stalk (secondary constriction).

Supplementary Figure S15. The proportion of false-positive variants that overlap with different genome stratifications across tumour purities. Four types of genome stratifications were examined for SNVs (A) and indels (B) called by ClairS and DeepSomatic in COLO829: segmental duplications, tandem repeats, short read low mappability regions, and abnormal GC% regions. The colour of tiles represents the proportion of variants that overlap with specific genome stratifications.

Supplementary Figure S16. The proportion of false-positive SVs that overlap with different genome stratifications in samples with 100% tumour purity. SVs specific to LRS are termed false-positive events compared to short reads. Genome stratifications were labelled on the x-axis with results from four tools listed on the horizontal panels for COLO829 (left) and HCC1937 (right). Bars in teal represent the SV proportions that overlap with specific genome stratifications.

Supplementary Figure S17. Germline leakage rate across tumour purity and sequencing depth. (A) The germline leakage rate in samples of the COLO829 and HCC1937 cell lines with tumour purity from 100% to 10% for ClairS and DeepSomatic. (B) The germline leakage rate in variant calls with different tumour and normal read depth combinations for ClairS and DeepSomatic in COLO829 and HCC1937. Only tumour samples with 100% tumour purity were shown in the plot.

Supplementary Figure S18. Benchmarking SV detection using the public COLO829 gold standard. (A) Heatmap showing the 68 gold standard SVs (x-axis) that were detected or missed by four tools across different tumour purities, with colours indicating the SV types. (B) Bar plots display the recall and precision rates calculated for each tool under different tumour purities, with colours indicating the FP, FN, or TP.

Supplementary Table S1. Comparison of LRS alignments between CHM13-T2T and GRCh38 human reference assemblies.

Supplementary Table S2. SNV detection of intersections and unions between ClairS and DeepSomatic across tumour purity.

Supplementary Table S3. The somatic SNV calling performance on chromosome 22 using ClairS and DeepSomatic.

Author contributions

Jia Zhang (Conceptualization [lead], Formal analysis [lead], Investigation [lead], Methodology [lead], Software [lead], Visualization [lead], Writing – original draft [lead]), Ho Yi Wong (Investigation [supporting], Resources [supporting], Writing – review & editing [supporting]), Lingchen Liu (Methodology [supporting], Writing – review & editing [supporting]), Lambros T. Koufariotis (Formal analysis [supporting], Resources [supporting], Writing – review & editing [supporting]), Scott Wood (Resources [supporting], Writing – review & editing [supporting]), Nadine Fitzpatrick (Investigation [supporting], Resources [supporting], Writing – review & editing [supporting]), Jenny Quiatchon (Investigation [supporting], Resources [supporting], Writing – review & editing [supporting]).

ing]], Paul Collins (Investigation [supporting], Resources [supporting], Supervision [supporting], Writing – review & editing [supporting]), John V. Pearson (Conceptualization [equal], Data curation [supporting], Funding acquisition [equal], Methodology [supporting], Resources [supporting], Supervision [equal], Writing – review & editing [supporting]) and Nicola Waddell (Conceptualization [equal], Data curation [supporting], Funding acquisition [equal], Methodology [supporting], Resources [supporting], Supervision [equal], Writing – review & editing [supporting])

Funding

This work was funded by a Cancer Council Queensland (CCQ) Accelerating Collaborative Cancer Research (AACR) Grant (000000027). N.W. is funded by the National Health and Medical Research Council of Australia (NHMRC) Investigator Grant (2018244). L.L. is supported by a University of Queensland Graduate School Scholarship, QIMR Berghofer PhD Top-Up Scholarship. This research was performed on QIMR Berghofer computing infrastructure supported by the Ian Potter Foundation and the Australian Cancer Research Foundation (ACRF) through the ACRF Centre for Optimised Cancer Therapy.

Data availability

All additional supporting data are available in the *GigaScience* repository, GigaDB [87]. The BAM files for cell line mixtures are available from the European Genome-Phenome Archive (EGA) under accession number EGAD00001015628 under study EGAS00001008107. Raw POD5 files are available on request.

Competing interests

J.V.P. and N.W. are co-founders of genomiQa. L.L. and N.W. were funded by Oxford Nanopore to present work from this study at meetings. The remaining authors declare that there are no competing interests.

References

- Alexandrov LB, Kim J, Haradhvala NJ, et al. The repertoire of mutational signatures in human cancer. *Nature*. 2020;**578**:94–101. <https://doi.org/10.1038/s41586-020-1943-3>.
- Hanahan D. Hallmarks of cancer: new dimensions. *Cancer Discov*. 2022;**12**:31–46. <https://doi.org/10.1158/2159-8290.CD-21-1059>.
- ICGC/TCGA. Pan-cancer analysis of whole genomes consortium. pan-cancer analysis of whole genomes. *Nature*. 2020;**578**:82–93. <https://doi.org/10.1038/s41586-020-1969-6>.
- Yokoyama A, Kakiuchi N, Yoshizato T, et al. Age-related remodelling of oesophageal epithelia by mutated cancer drivers. *Nature*. 2019;**565**:312–17. <https://doi.org/10.1038/s41586-018-0811-x>.
- Ramarao-Milne P, Kondrashova O, Patch A-M, et al. Comparison of actionable events detected in cancer genomes by whole-genome sequencing, in silico whole-exome and mutation panels. *ESMO Open*. 2022;**7**:100540. <https://doi.org/10.1016/j.esmoop.2022.100540>.
- Lee JK, Sivakumar S, Schrock AB, et al. Comprehensive pan-cancer genomic landscape of KRAS altered cancers and real-world outcomes in solid tumors. *NPJ Precis Onc*. 2022;**6**:91. <https://doi.org/10.1038/s41698-022-00334-z>.
- Chakravarty D, Solit DB. Clinical cancer genomic profiling. *Nat Rev Genet*. 2021;**22**:483–501. <https://doi.org/10.1038/s41576-021-00338-8>.
- Dwarshuis N, Kalra D, McDaniel J, et al. The GIAB genomic stratifications resource for human reference genomes. *Nat Commun*. 2024;**15**:9029. <https://doi.org/10.1038/s41467-024-53260-y>.
- Ermini L, Driguez P. The application of long-read sequencing to cancer. *Cancers*. 2024;**16**:1275. <https://doi.org/10.3390/cancers16071275>.
- Mahmoud M, Gobet N, Cruz-Dávalos DI, et al. Structural variant calling: the long and the short of it. *Genome Biol*. 2019;**20**:246. <https://doi.org/10.1186/s13059-019-1828-7>.
- Rhoads A, Au KF. PacBio sequencing and its applications. *Genomics Proteomics Bioinformatics*. 2015;**13**:278–89. <https://doi.org/10.1016/j.gpb.2015.08.002>.
- Wang Y, Zhao Y, Bollas A, et al. Nanopore sequencing technology, bioinformatics and applications. *Nat Biotechnol*. 2021;**39**:1348–1365. <https://doi.org/10.1038/s41587-021-01108-x>.
- Huddleston J, Ranade S, Malig M, et al. Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Res*. 2014;**24**:688–96. <https://doi.org/10.1101/gr.168450.113>.
- Nurk S, Koren S, Rhie A, et al. The complete sequence of a human genome. *Science*. 2022;**376**:44–53. <https://doi.org/10.1126/science.abj6987>.
- Kuleshov V, Xie D, Chen R, et al. Whole-genome haplotyping using long reads and statistical methods. *Nat Biotechnol*. 2014;**32**:261–66. <https://doi.org/10.1038/nbt.2833>.
- Sakamoto Y, Sereewattanawoot S, Suzuki A. A new era of long-read sequencing for cancer genomics. *J Hum Genet*. 2020;**65**:3–10. <https://doi.org/10.1038/s10038-019-0658-5>.
- Nattestad M, Goodwin S, Ng K, et al. Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line. *Genome Res*. 2018;**28**:1126–35. <https://doi.org/10.1101/gr.231100.117>.
- Cui X, Lin Q, Chen M, et al. Long-read sequencing unveils novel somatic variants and methylation patterns in the genetic information system of early lung cancer. *Comput Biol Med*. 2024;**171**:108174. <https://doi.org/10.1016/j.compbiomed.2024.108174>.
- Xu L, Wang X, Lu X, et al. Long-read sequencing identifies novel structural variations in colorectal cancer. *PLoS Genet*. 2023;**19**:e1010514. <https://doi.org/10.1371/journal.pgen.1010514>.
- O'Neill K, Pleasance E, Fan J, et al. Long-read sequencing of an advanced cancer cohort resolves rearrangements, unravels haplotypes, and reveals methylation landscapes. *Cell Genomics*. 2024;**4**:100674. <https://doi.org/10.1016/j.xgen.2024.100674>.
- Brändl B, Steiger M, Kubelt C, et al. Rapid brain tumor classification from sparse epigenomic data. *Nat Med*. 2025;**31**:840–48. <https://doi.org/10.1038/s41591-024-03435-3>.

22. Patel A, Göbel K, Ille S, et al. Prospective, multicenter validation of a platform for rapid molecular profiling of central nervous system tumors. *Nat Med*. 2025;**31**:1567–77. <https://doi.org/10.1038/s41591-025-03562-5>.
23. Amarasinghe SL, Ritchie ME, Gouil Q. Long-read-tools.Org: an interactive catalogue of analysis methods for long-read sequencing data. *GigaScience*. 2021;**10**:giab003. <https://doi.org/10.1093/gigascience/giab003>.
24. Jain M, Koren S, Miga KH, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol*. 2018;**36**:338–45. <https://doi.org/10.1038/nbt.4060>.
25. Sedlazeck FJ, Lee H, Darby CA, et al. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat Rev Genet*. 2018;**19**:329–46. <https://doi.org/10.1038/s41576-018-0003-4>.
26. LoTempio J, Delot E, Vilain E. Benchmarking long-read genome sequence alignment tools for human genomics applications. *PeerJ*. 2023;**11**:e16515. <https://doi.org/10.7717/peerj.16515>.
27. Liu L, Zhang J, Wood S, et al. Performance of somatic structural variant calling in lung cancer using Oxford Nanopore sequencing technology. *BMC Genomics*. 2024;**25**:898. <https://doi.org/10.1186/s12864-024-10792-3>.
28. Sigurpalsdottir BD, Stefansson OA, Holley G, et al. A comparison of methods for detecting DNA methylation from long-read sequencing of human genomes. *Genome Biol*. 2024;**25**:69. <https://doi.org/10.1186/s13059-024-03207-9>.
29. Dong X, Du MRM, Gouil Q, et al. Benchmarking long-read RNA-sequencing analysis tools using in silico mixtures. *Nat Methods*. 2023;**20**:1810–21. <https://doi.org/10.1038/s41592-023-02026-3>.
30. Liu Y, Rosikiewicz W, Pan Z, et al. DNA methylation-calling tools for Oxford Nanopore sequencing: a survey and human epigenome-wide evaluation. *Genome Biol*. 2021;**22**:295. <https://doi.org/10.1186/s13059-021-02510-z>.
31. Park J, Cook DE, Chang P-C, et al. Accurate somatic small variant discovery for multiple sequencing technologies with DeepSomatic. *Nat Biotechnol*. 2025. <https://doi.org/10.1038/s41587-025-02839-x>.
32. Chen L, Zheng Z, Su J, et al. ClairS-TO: a deep-learning method for long-read tumor-only somatic small variant calling. *Nat Commun*. 2025;**16**:9630. <https://doi.org/10.1038/s41467-025-64547-z>.
33. Keskus A, Bryant A, Ahmad T, et al. Severus detects somatic structural variation and complex rearrangements in cancer genomes using long-read sequencing. *Nat Biotechnol*. 2026;**44**: 247–257. <https://doi.org/10.1038/s41587-025-02618-8>.
34. Elrick H, Sauer CM, Trevers K, et al. SAVANA: reliable analysis of somatic structural variants and copy number aberrations using long-read sequencing. *Nat Methods*. 2025;**22**:1436–46. <https://doi.org/10.1038/s41592-025-02708-0>.
35. Griffith M, Miller CA, Griffith OL, et al. Optimizing cancer genome sequencing and analysis. *Cell Syst*. 2015;**1**:210–23. <https://doi.org/10.1016/j.cels.2015.08.015>.
36. Xiao W, Ren L, Chen Z, et al. Toward best practice in cancer mutation detection with whole-genome and whole-exome sequencing. *Nat Biotechnol*. 2021;**39**:1141–50. <https://doi.org/10.1038/s41587-021-00994-5>.
37. Yu T, Huang Q, Zhao X, et al. Tumour purity as an underlying key factor in tumour mutation detection in colorectal cancer. *Clinical & Translational Med*. 2023;**13**:e1252. <https://doi.org/10.1002/ctm2.1252>.
38. Haider S, Tyekucheva S, Prandi D, et al. Systematic assessment of tumor purity and its clinical implications. *JCO Precis Oncol*. 2020;**4**:995–1005. <https://doi.org/10.1200/PO.20.00016>.
39. Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. *Nat Commun*. 2015;**6**:8971. <https://doi.org/10.1038/ncomms9971>.
40. Mwenifumbo JC, Marra MA. Cancer genome-sequencing study design. *Nat Rev Genet*. 2013;**14**:321–32. <https://doi.org/10.1038/nrg3445>.
41. Sendorek DH, Caloian C, Ellrott K, et al. Germline contamination and leakage in whole genome somatic single nucleotide variant detection. *BMC Bioinf*. 2018;**19**:28. <https://doi.org/10.1186/s12859-018-2046-0>.
42. Chen S, Francioli LC, Goodrich JK, et al. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature*. 2024;**625**:92–100. <https://doi.org/10.1038/s41586-023-06045-0>.
43. Espejo Valle-Inclan J, Besseling NJM, de Bruijn E, et al. A multi-platform reference for somatic structural variation detection. *Cell Genomics*. 2022;**2**:100139. <https://doi.org/10.1016/j.xgen.2022.100139>.
44. Yuen ZW-S, Srivastava A, Daniel R, et al. Systematic benchmarking of tools for CpG methylation detection from nanopore sequencing. *Nat Commun*. 2021;**12**:3438. <https://doi.org/10.1038/s41467-021-23778-6>.
45. Pardo-Palacios FJ, Wang D, Reese F, et al. Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. *Nat Methods*. 2024;**21**:1349–63. <https://doi.org/10.1038/s41592-024-02298-3>.
46. Alioto TS, Buchhalter I, Derdak S, et al. A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nat Commun*. 2015;**6**:10001. <https://doi.org/10.1038/ncomms10001>.
47. Poplin R, Chang P-C, Alexander D, et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol*. 2018;**36**:983–87. <https://doi.org/10.1038/nbt.4235>.
48. Le MK, Qin Q, colorSV LH: Long-range somatic structural variation calling from matched tumor-normal co-assembly graphs. *Genomics Proteomics Bioinformatics*. 2025;qzaf082 <https://doi.org/10.1093/gpbjnl/qzaf082>.
49. De L, Cerda GY, Landis JB, et al. Balancing read length and sequencing depth: Optimizing Nanopore long-read sequencing for monocots with an emphasis on the Liliales. *Appl Plant Sci*. 2023;**11**:e11524. <https://doi.org/10.1002/aps3.11524>.
50. Kosugi S, Terao C. Comparative evaluation of SNVs, indels, and structural variations detected with short- and long-read sequencing data. *Hum Genome Var*. 2024;**11**:18. <https://doi.org/10.1038/s41439-024-00276-x>.
51. Choo Z-N, Behr JM, Deshpande A, et al. Most large structural variants in cancer genomes can be detected without long reads. *Nat Genet*. 2023;**55**:2139–48. <https://doi.org/10.1038/s41588-023-01540-6>.
52. Kolmogorov M, Billingsley KJ, Mastoras M, et al. Scalable Nanopore sequencing of human genomes provides a com-

- prehensive view of haplotype-resolved variation and methylation. *Nat Methods*. 2023;**20**:1483–92. <https://doi.org/10.1038/s41592-023-01993-x>.
53. Nyaga DM, Tsai P, Gebbie C, et al. Benchmarking nanopore sequencing and rapid genomics feasibility: validation at a quaternary hospital in New Zealand. *npj Genom Med*. 2024;**9**:57. <https://doi.org/10.1038/s41525-024-00445-5>.
 54. Olson ND, Wagner J, Dwarshuis N, et al. Variant calling and benchmarking in an era of complete human genome sequences. *Nat Rev Genet*. 2023;**24**:464–83. <https://doi.org/10.1038/s41576-023-00590-0>.
 55. Paulin LF, Fan J, O'Neill K, et al. The benefit of a complete reference genome for cancer structural variant analysis. *medRxiv*. 2024. <https://doi.org/10.1101/2024.03.15.24304369>.
 56. Ameur A, Bunikis I, Ek M, et al. Long-read sequencing and optical mapping generates near T2T assemblies that resolves a centromeric translocation. *Sci Rep*. 2024;**14**:9000. <https://doi.org/10.1038/s41598-024-59683-3>.
 57. Li H, Bloom JM, Farjoun Y, et al. A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nat Methods*. 2018;**15**:595–97. <https://doi.org/10.1038/s41592-018-0054-7>.
 58. Yang C, Chu J, Warren RL, et al. NanoSim: nanopore sequence read simulator based on statistical characterization. *GigaScience*. 2017;**6**:1–6. <https://doi.org/10.1093/gigascience/gix010>.
 59. Ono Y, Asai K, Hamada M. PBSIM2: a simulator for long-read sequencers with a novel generative model of quality scores. *Bioinformatics*. 2021;**37**:589–95. <https://doi.org/10.1093/bioinformatics/btaa835>.
 60. Craig DW, Nasser S, Corbett R, et al. A somatic reference standard for cancer genome sequencing. *Sci Rep*. 2016;**6**:24607. <https://doi.org/10.1038/srep24607>.
 61. Ewing AD, Houlahan KE, Hu Y, et al. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nat Methods*. 2015;**12**:623–30. <https://doi.org/10.1038/nmeth.3407>.
 62. English AC, Dolzhenko E, Ziaei Jam H, et al. Analysis and benchmarking of small and large genomic variants across tandem repeats. *Nat Biotechnol*. 2025;**43**:431–42. <https://doi.org/10.1038/s41587-024-02225-z>.
 63. Dorado: Oxford Nanopore's Basecaller. <https://github.com/nanoporetech/dorado/>. (Accessed Dec 2023).
 64. Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;**25**:2078–79. <https://doi.org/10.1093/bioinformatics/btp352>.
 65. Li H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics*. 2021;**37**:4572–74. <https://doi.org/10.1093/bioinformatics/btab705>.
 66. Tarasov A, Vilella AJ, Cuppen E, et al. Sambamba: fast processing of NGS alignment formats. *Bioinformatics*. 2015;**31**:2032–34. <https://doi.org/10.1093/bioinformatics/btv098>.
 67. Pedersen BS, Quinlan AR. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics*. 2018;**34**:867–68. <https://doi.org/10.1093/bioinformatics/btx699>.
 68. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. *GigaScience*. 2021;**10**:giab008. <https://doi.org/10.1093/gigascience/giab008>.
 69. Rausch T, Zichner T, Schlattl A, et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*. 2012;**28**:i333–i339. <https://doi.org/10.1093/bioinformatics/bts378>.
 70. Shiraishi Y, Koya J, Chiba K, et al. Precise characterization of somatic complex structural variations from tumor/control paired long-read sequencing data with nanomonsv. *Nucleic Acids Res*. 2023;**51**:e74. <https://doi.org/10.1093/nar/gkad526>.
 71. Martin M, Ebert P, Marschall T. Read-based phasing and analysis of phased variants with WhatsHap. *Methods Mol Biol*. 2023;**2590**:127–38. https://doi.org/10.1007/978-1-0716-2819-5_8.
 72. Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr Protoc Bioinformatics*. 2013;**43**:11.10.1–11.10.33. <https://doi.org/10.1002/0471250953.bi1110s43>.
 73. Kassahn KS, Holmes O, Nones K, et al. Somatic point mutation calling in low cellularity tumors. *PLoS One*. 2013;**8**:e74380. <https://doi.org/10.1371/journal.pone.0074380>.
 74. Cameron DL, Baber J, Shale C, et al. GRIDSS2: comprehensive characterisation of somatic structural variation using single breakend variants and structural variant phasing. *Genome Biol*. 2021;**22**:202. <https://doi.org/10.1186/s13059-021-02423-x>.
 75. Layer RM, Chiang C, Quinlan AR, et al. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol*. 2014;**15**:R84. <https://doi.org/10.1186/gb-2014-15-6-r84>.
 76. Kirsche M, Prabhu G, Sherman R, et al. Jasmine and Iris: population-scale structural variant comparison and analysis. *Nat Methods*. 2023;**20**:408–17. <https://doi.org/10.1038/s41592-022-01753-3>.
 77. Aryee MJ, Jaffe AE, Corrada-Bravo H, et al. Minfi: a flexible and comprehensive bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*. 2014;**30**:1363–69. <https://doi.org/10.1093/bioinformatics/btu049>.
 78. Lawrence M, Huber W, Pagès H, et al. Software for computing and annotating genomic ranges. *PLoS Comput Biol*. 2013;**9**:e1003118. <https://doi.org/10.1371/journal.pcbi.1003118>.
 79. Lawrence M, Gentleman R, Carey V. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics*. 2009;**25**:1841–42. <https://doi.org/10.1093/bioinformatics/btp328>.
 80. Nassar LR, Barber GP, Benet-Pagès A, et al. The UCSC Genome Browser database: 2023 update. *Nucleic Acids Res*. 2023;**51**:D1188–D95. <https://doi.org/10.1093/nar/gkac1072>.
 81. Obenchain V, Lawrence M, Carey V, et al. VariantAnnotation: a bioconductor package for exploration and annotation of genetic variants. *Bioinformatics*. 2014;**30**:2076–78. <https://doi.org/10.1093/bioinformatics/btu168>.
 82. Perez G, Barber GP, Benet-Pagès A, et al. The UCSC Genome Browser database: 2025 update. *Nucleic Acids Res*. 2025;**53**:D1243–49. <https://doi.org/10.1093/nar/gkae974>.
 83. Rosenthal R, McGranahan N, Herrero J, et al. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma

- evolution. *Genome Biol.* 2016;**17**:31. <https://doi.org/10.1186/s13059-016-0893-4>.
84. Wickham H, Averick M, Bryan J, et al. Welcome to the tidyverse. *JOSS.* 2019;**4**:1686. <https://doi.org/10.21105/joss.01686>.
85. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer, 2016. <https://doi.org/10.1007/978-3-319-24277-4>.
86. van den Brand T. ggh4x: Hacks for “ggplot2.” R package version 0.3.1.9000. 2025. <https://github.com/teunbrand/ggh4x>. Accessed Jan 2024.
87. Zhang J, Wong HY, Liu L, et al. Supporting data for “cancer genome standards for long-read sequencing using cancer cell line mixtures”. GigaScience Database. 2026. <https://doi.org/10.5524/102811>.