Scalable and cloud-enabled analysis of long read sequencing data

Shishir Reddy

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2022

Committee:

Ka Yee Yeung

Ling-Hong Hung

Cecilia Yeung

Olga Sala-Torra

Program Authorized to Offer Degree:

Computer Science and Systems

University of Washington

Abstract

Scalable and cloud-enabled analysis of long read sequencing data

Shishir Reddy

Chair of the Supervisory Committee:
Ka Yee Yeung
School of Engineering & Technology

Long-read sequencing has great promise in enabling portable, rapid molecular-assisted diagnoses. Applications of long-read sequencing include improved prognosis of critically ill patients through variant detection along with rapid genetic diagnoses. A key challenge in democratizing long-read sequencing technology in the biomedical and clinical community is the lack of graphical bioinformatics software tools which can efficiently process the raw data, support graphical output and interactive visualizations for interpretations of results. Another obstacle is that high performance software tools for long-read sequencing data analyses often leverage graphics processing units (GPU), which is challenging and time-consuming to configure, especially on the cloud.

Many solutions can be explored in long-read sequencing including the addition of graphical bioinformatics software tools, hardware acceleration such as Graphics Processing Units (GPUs), or optimization with Tensor Processing Units (TPUs). Long-read sequencing workflows for diagnosis involve several steps that can be hardware-accelerated and optimized using various processing methods. Optimizing long-read sequencing workflows through hardware-acceleration can reduce turnaround times of diagnoses from days to hours. Our goal is to create and optimize long-read sequencing workflows to build rapid, cost-effective solutions for cancer detection and diagnosis on the cloud.

This thesis introduces two containerized, hardware-accelerated long-read sequencing anal-

ysis workflows for fusion analysis and variant-calling. The fusion analysis workflow introduces a fusion finding tool – the Biodepot Fusion Finder (BFF) – capable of rapidly detecting fusions and calculating sample enrichment. This fusion workflow is benchmarked for accuracy and compared to the fusion finding software LongGF on cell-line and patient samples of nanopore data. The variant-calling workflow uses PEPPER-Margin-Deepvariant to call structural variants in a cloud-based GPU-enabled environment. This workflow is benchmarked for accuracy between GPU and CPU versions of the variant-calling software for better visibility in which specific stages of the pipeline benefit from hardware acceleration.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to thank Dr. Cecilia Yeung and Dr. Olga Sala Torra for introducing me to nanopore sequencing and for providing data and guidance vital to this thesis. I would also like to thank Dr. Ka Yee Yeung and Dr. Ling-Hong Hung for their incredible support and guidance throughout my academic career here at the University of Washington Tacoma.

Chapter 1

INTRODUCTION

1.1 Background

Advances in molecular diagnosis have enabled detection of specific driver mutations that can be essential for prognosis, monitoring, and targeted therapy [15, 31, 30]. Examples of such "precision medicine" include the BCR-ABL fusion gene in chronic myeloid leukemia (CML), PML-RARA fusions in acute promyelocytic leukemia (APL) and FLT3 mutations in acute myeloid leukemia (AML) [20, 12, 16, 6]. Potentially targetable mutations are also found in solid tumors, such as renal cell carcinoma [29, 26, 17]. To capitalize on the potential of precision medicine, accelerated analysis of sequencing data is needed to improve the potential of molecular-assisted cancer diagnoses [27]. We address this using long-read sequencing technology, such as Oxford Nanopore Technologies (ONT), which generates continuous sequences up to a few megabases in length and can directly sequence DNA, resulting in much shorter turnaround times than next generation sequencing (NGS) [11, 9, 7].

Computational methods and software tools tailored for long-read sequencing data are essential to enable the use of this emerging and promising technology [24, 3]. Recurrent gene fusions are common drivers of disease pathophysiology in leukemias. Identification of these structural variants helps stratify disease by risk and assists with therapy choice. Current fusion detection methods require long turnaround time (7-10 days) or advance knowledge of the genes involved in the fusions [23]. Thus, rapid sequencing technology such as the nanopore sequencing offered by ONT is a good fit for fusion detection and variant-calling applications. In nanopore sequencing, electrical current alterations are recorded as different bases traverse the pore opening. Basecalling, which translates the signal (stored as fast5 files) into a sequence of base pairs is the key step determining accuracy of the sequencing

experiment. Basecalling is computationally expensive and a rate-limiting step in the analysis of nanopore data. Deep learning neural network models have been applied to basecalling to increase the accuracy [28]. With standard CPU processing, these methods are prohibitively slow and require large numbers of computational cores operating in parallel to be practical. Graphics processing units (GPUs) can be used to accelerate the analysis but require specialized hardware and software.

This hardware and software acceleration can be extended past the basecalling step, creating a need for optimized nanopore workflows that are easy to use and benefit from hardware-acceleration in compute-intensive steps. After the nanopore data is basecalled, the reads are aligned to the human genome with Minimap2 [13] and can be visualized with the Integrated Genome Viewer (IGV) [22]. The aligned reads can then be analyzed for mutations with fusion detection or variant calling steps depending on the types of mutations located within the sample. The variant calling step, like basecalling, has potential to benefit greatly from hardware-acceleration as most modern variant callers utilize deep learning models.

1.2 Contributions

In this thesis project, we aim to expand support for more nanopore workflows, optimize the performance of these workflows, and collaborate with external sources to produce cancer diagnostics. These contributions will be accomplished through the following tasks

- Additions of variant calling and fusion-finding software to existing nanopore workflows in order to support a broader range of detection.

- The runtime and accuracies of existing workflows will be optimized through parameter tuning, model selection, hardware acceleration.

- Collaboration with Fred Hutchinson Cancer Center for cancer diagnostics on patient and cell-line data.

Chapter 2

RELATED WORK

2.1 Existing Software Tools and Workflows

2.1.1 Basecalling: Methods and Tools

We focus on two basecalling software: Guppy and Bonito. Guppy [2] is a proprietary data processing toolkit that contains the Oxford Nanopore Technologies' basecalling algorithms and several post-processing features. A selection of configuration files allows basecalling of DNA and RNA libraries made with Oxford Nanopore Technologies' current sequencing kits. Guppy basecalling models are based on Recurrent Neural Networks (RNN) and benefit greatly from GPU-acceleration. Bonito [1] is a research-grade basecaller created by ONT that is more accessible. Bonito relies on a convolutional layer for training and inference and could be accelerated using specialized hardware (such as GPU).

2.1.2 Variant Calling: Methods and Tools

Variant callers identify true changes in DNA sequence from a long string of noisy data. The primary variant calling software used in this workflow is PEPPER-Margin-Deepvariant [25]. This variant caller was a top performer in the PrecisionFDA Truth Challenge V2 [5], where variant calling pipeline performance was assessed on a common reference. This variant caller can be split into five stages and uses three models: Recurrent neural networks (RNN), Hidden Markov Model (HMM), and a deep convolutional neural network (CNN). The RNN and CNN stages of this variant caller have potential to benefit greatly from GPU-acceleration, while the CNN has further potential runtime improvements with TPU-acceleration.

2.1.3 Preliminary Results by Reddy et al.

The effects of GPU acceleration in long-read sequencing workflows has been previously benchmarked against traditional CPU methods. I recently published a long-read sequencing workflow supported by an Amazon Machine Image (AMI) with software and drivers pre-installed for GPU computing on the cloud [21]. This paper demonstrates that computational methods and software tools tailored for long-read sequencing data are essential to enable use of long-read sequencing workflows. Two crucial steps in traditional long-read sequencing pipelines are supported by deep neural networks which benefit greatly from GPU hardware-acceleration. Benchmarks from GPU and CPU workflows show a 29x speedup in GPU computing with a 93x reduction in cloud computing costs [21]. This preliminary workflow is shown in Figure 2.1.

2.2 Applications of Long-Read Sequencing Workflows

2.2.1 Structural Variant Applications by Amarasinghe et al.

Amarasinghe et al. explore the variety of long-read sequencing tools available and detail their applications in detecting structural variants. Short reads perform well for identifying single nucleotide variants (SNVs) and small insertion and deletions (indels), but they are inaccurate in detecting larger sequence changes [4]. Structural variants that affect greater than 50 base pairs are better suited for long-read sequencing [4]. At the time of this publication's release, long-read structural variant callers could not be reliably assessed due to the lack of structural variants in the benchmark data sets. This problem has since been resolved with the rapid development of long-read sequencing technologies and variant callers.

2.2.2 Clinical Diagnosis in a Critical Care Setting by Gorzynski et al.

In the correspondence *Ultrarapid Nanopore Genome Sequencing in a Critical Care Setting* recently published in the prestigious New England Journal of Medicine, Gorzynski et al. demonstrate the applications of long-read sequencing workflows in prognosis improvement,

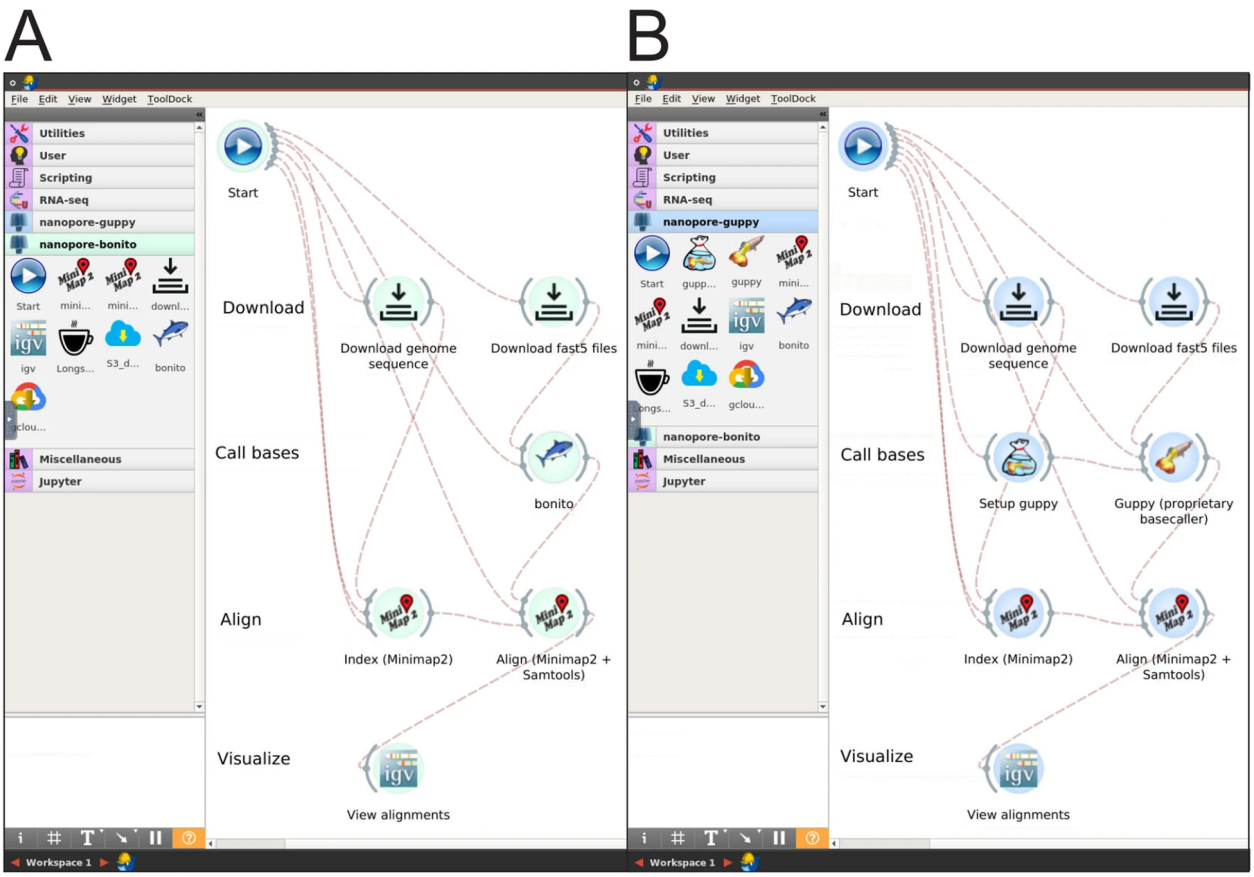


Figure 2.1: Screenshots of our interactive GPU workflow which uses the Biodepot-workflow-builder platform. Panel A is a screenshot of the workflow using the open-source Bonito basecaller. Panel B is a screenshot of the workflow using the proprietary Guppy basecaller. Both basecallers use GPUs.

rapid diagnosis, and cost reduction for critically ill patients. A long-read sequencing workflow is showcased and advertised to permit clinical diagnosis within 8 hours of sample collection. The major optimization made in this correspondence that is relevant to this thesis is the real-time basecalling and alignment structure of the workflow that runs several steps of the pipeline in parallel [8].

2.3 Comparison

Reddy et al. and Gorzynski et al. demonstrate the potential offered by long-read sequencing in producing rapid, molecular-assisted diagnoses. Efficient workflows are constructed by both groups with the goal of reducing turnaround time to hours for same-day diagnosis. Reddy et al. provide a GPU-accelerated implementation with a focus on reduction in runtime during the basecalling step. Gorzynski et al. offer runtime optimizations with parallel execution of certain steps in the workflow and a more robust example with the inclusion of a GPU-accelerated variant calling step. Aspects of both implementations can be combined to achieve a further optimized nanopore sequencing workflow. Basecalling and variant calling steps can also be optimized with the introduction of hardware-acceleration targeted towards deep learning compute-intensive steps.

Chapter 3

EXPERIMENT DESIGN

3.1 Overview

Our goal is to create efficient nanopore sequencing workflows optimized with hardware-acceleration that support variant calling and fusion detection applications. Optimization of these workflows are done through benchmarking individual compute-intensive steps with differing hardware, software, and adjustable hyperparameters. The data used for testing and comparison in this experiment has been provided by Dr. Cecilia Yeung, Dr. Olga Sala-Torra, and Dr. Jerald Radich from Fred Hutchinson Cancer Center. This data is used to measure the overall accuracy of the workflow while also maintaining a standard for overall runtime.

3.2 Data

The datasets used in this project can be split into two major categories: patient and cell-line. Cell-line data stems from a well-known reference of the human genome and contains mutations widely documented by researchers. Patient data refers to data obtained directly from patients with unique mutation locations not widely-researched or documented. Cell-line data is easily-navigated as mutations are well-documented and a ground-truth set is known. Patient data involves many unknowns both in number of mutations and locations.

The cell-line data used in this project contains fusion mutations between the PML and RARA genes as well as between the BCR and ABL1 genes. Both patient and cell-line data were provided by committee members. We can reliably detect fusion genes from DNA sequences using cell lines (NB4, K562, ME1, and MV411) with known fusion genes. The patient data provided covers a wider variety of fusions including BCR-ABL, PML-RARA, CBFB-MYH11, and KMT2A-AF4.

For variant calling benchmarks, Next-Generation Sequencing (NGS) data is used with Deepvariant [18] and nanopore data is used with PEPPER-Margin-Deepvariant. The NGS dataset used is the same as the HG003 test dataset provided in the PrecisionFDA Truth Challenge V2 [5]. DNA extracted from a single large batch of cells for the three genomes (son - HG002, father – HG003, and mother - HG004) is publicly available in National Institute of Standards and Technology Reference Materials 8391 (HG002) and 8392 (HG002-HG004). The Genome in a Bottle Consortium selected these genomes for characterization as they are a trio from the Personal Genome Project that has a broader consent permitting commercial redistribution and recontacting participants for further sample collection. Chromosome 20 of the HG003 sample was used here for benchmarking Deepvariant. The nanopore patient sample AML1 was used for benchmarking PEPPER-Margin-Deepvariant. This sample contains a PML-RARA fusion and 5.47 Gigabases of data.

3.3 Methods

The experiments for this thesis take part in several steps. First, a baseline workflow has been created without hardware-acceleration to serve as a standard for comparison. Next, each compute-intensive step of the workflow has been optimized through several means: hardware acceleration, hyperparameter adjustment, and model selection. Afterwards, additional software is introduced and benchmarked for fusion detection and variant calling. Finally, the resulting optimized workflows are containerized and packaged for simple modification and deployment.

3.3.1 Baseline

The baseline workflow for this project will serve as a reference during benchmarking of workflow optimizations. This workflow contains the proprietary basecaller Guppy in the basecalling step followed by Minimap2 [13] for alignment, and finally the Integrative Genomics Viewer (IGV) [22] for visualization. The baseline results for basecalling are measured using an AWS EC2 c5d18xlarge instance with 72 vCPUs, 72 threads/basecaller, and 1 basecaller.

This is used to measure the performance of basecalling on all datasets without any additional hardware-acceleration.

This baseline workflow will be compared to a hardware-accelerated version using an AWS g4dn.4xlarge virtual machine instance with a NVIDIA Tesla T4 GPU. This workflow is also benchmarked locally on a laptop with a GeForce RTX 2060 GPU. All benchmark experiments on AWS were based on 4 runs [21].

3.3.2 Fusion Detection

The first workflow developed focuses on rapid detection of myeloid neoplasm fusions [23]. This workflow adds two widgets for fusion detection: Biodepot-Fusion-Finder [23] and LongGF [14]. LongGF, a current state-of-the-art fusion detection tool for long-read sequencing data, struggles to detect noisier fusions. We created the fusion finding tool Biodepot-Fusion-Finder as an alternative for fusion detection to allow for larger gaps in fusion reads and better support of noisier data. Along with the option of fusion detection, this workflow also supports the calculation of a sample's enrichment through BFF, allowing users to reliably determine the accuracy and confidence within any specific sample. Two forms of sample enrichment are calculated: fusion and on-target. Fusion enrichment refers to the number of targeted fusion reads that contain the breakpoints over the mean genome coverage. On-target encrichment calculates the number of reads from the specified guide cut point to the fusion breakpoint over the mean genome coverage.

3.3.3 Variant Calling

The second workflow focuses on a complete, haplotype-aware variant calling pipeline starting with basecalling and ending with variant call format output. This workflow adds a widget for the variant caller PEPPER-Margin-Deepvariant [25] to the baseline workflow. This variant caller is tailored towards nanopore data and is trained on data produced by the latest versions of Oxford Nanopore Technology's proprietary basecaller Guppy.

3.4 Benchmarking

3.4.1 Accuracy

Accuracy is the primary benchmark for the fusion detection workflow. This benchmark focuses on the amount of fusions found within a sample between Biodepot-Fusion-Finder and LongGF independently. Each software will be tested individually on 6 cell-line samples and 14 patient samples. More fusions detected indicates a higher accuracy for fusion detection. The ground truth value of fusions detected for each sample has been confirmed manually through IGV (Integrated Genomics Viewer) [22].

For each individual sample, sequencing metrics including quality scores and timestamps are obtained from the sequencing summary text file obtained as an output of base calling using Guppy. Detected fusions are then acquired from the Breakpoint Finder as well as LongGF.

3.4.2 Runtime

Runtime (or execution time) serves as the primary benchmark for variant-calling. The runtime is measured across each step of the variant-callers Deepvariant and PEPPER-Margin-Deepvariant. The Deepvariant runtime benchmarks are used to demonstrate the potential of GPU acceleration within the inference step of the variant-caller. These runtime benchmarks are then extended to each individual step of PEPPER-Margin-Deepvariant, which is tailored towards long-read sequencing nanopore data. As this focus of this thesis is cloud-enabled analysis of long-read sequencing data, the final variant-calling workflow contains a containerized version of PEPPER-Margin-Deepvariant with hardware acceleration enabled. For both the GPU and CPU benchmarks, an AWS EC2 g4dn.4xlarge instance with 16 vCPUs, 64 GiB Memory, and a NVIDIA Tesla T4 GPU is used.

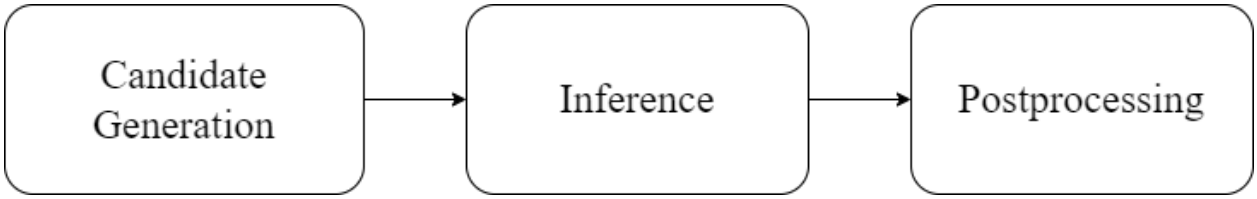


Figure 3.1: DeepVariant Stages.

Algorithms

Deepvariant's inference pipeline is split into three steps: candidate variant generation, variant prediction/inference, and postprocessing. See the flowchart in Figure 3.1. Candidate variant generation and postprocessing steps are CPU intensive and do not utilize a GPU even if one is provided. Variant prediction is GPU intensive and can potentially be the bottleneck if there are a large number of reads spread out over multiple chromosomes.

PEPPER-Margin-Deepvariant adds several steps to the variant calling pipeline. First, candidate variants are generated through the PEPPER-Candidate step using a recurrent neural network (RNN) [25]. Next, in the Margin-Phase step the candidate variants generated by PEPPER are haplotagged using a hidden Markov Model (HMM). Deepvariant is then run on candidate SNPs and INDELs separately. The results are finally merged into a resulting VCF (Variant Call Format) file. These stages are summarized in Figure 3.2.

3.5 Containerization and Deployment

After constructing optimized workflows extended to include variant calling and fusion detection software, the resulting long-read nanopore sequencing pipelines are containerized and deployed using the Biodepot-workflow-builder (Bwb) [10]. This provides modular and easy-to-use graphical interface for reproducible execution, customization and interactive visualization of the nanopore workflows.

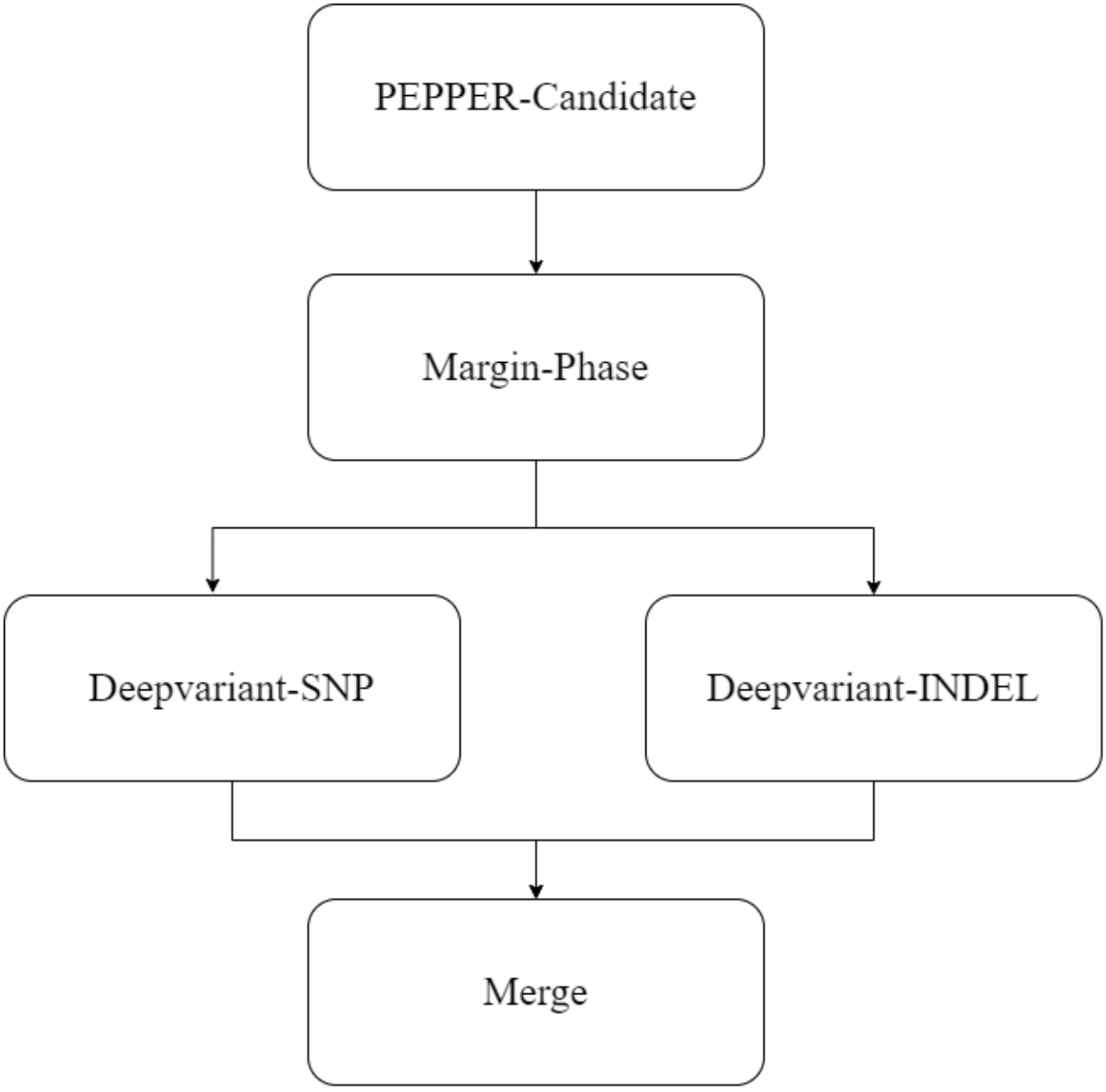


Figure 3.2: PEPPER-Margin-Deepvariant Stages.

Chapter 4

RESULTS AND DISCUSSION

4.1 Basecalling Benchmarks

We observed that Guppy GPU achieved the fastest average time at 88.9 s (1.5 min) with standard error of 1.2 s over 4 repeated measurements. Guppy CPU achieved the slowest average time at 2551.8 s (42.5 min) with standard error of 22.4 s. The 29x speedup is computed by comparing the average runtime (in seconds) of Guppy CPU to Guppy GPU (2551.8/88.9=28.7) [21].

Table 4.1: Comparison of Guppy CPU and GPU.

Basecaller	Cloud/Local	Average Runtime (seconds)	Standard Error (seconds)
Guppy CPU	AWS c5d18xlarge	2551.8	22.4
Guppy GPU	AWS g4dn.4xlarge	88.9	1.2
Guppy GPU	Laptop	135.3	0.6

4.2 Fusion Detection

4.2.1 Sample sequencing and enrichment

Details of the sample sequencing and enrichment metrics are included in Table 4.2. A range of 0.04 Gb – 5.47 gigabases of sequencing data was generated for each sample for an average mean coverage of the human genome of 0.32-fold (range: 0.01 – 1.66) [23].

4.2.2 Comparison of fusion detection tools

A comparison of the bioinformatic workflows for data analysis using different fusion detection widgets LongGF vs Biodepot Fusion Finder (BFF) was conducted; the specific workflow is demonstrated in Figure 4.1.

Along with detected fusions, BFF also computes the fusion enrichment and on target enrichment statistics; these are summarized in Table 4.2. In most cell line and primary specimens, LongGF shows a particular challenge in the detection of BCR-ABL1 and does not detect all fusion reads that are identified by BFF.

4.3 Variant Calling

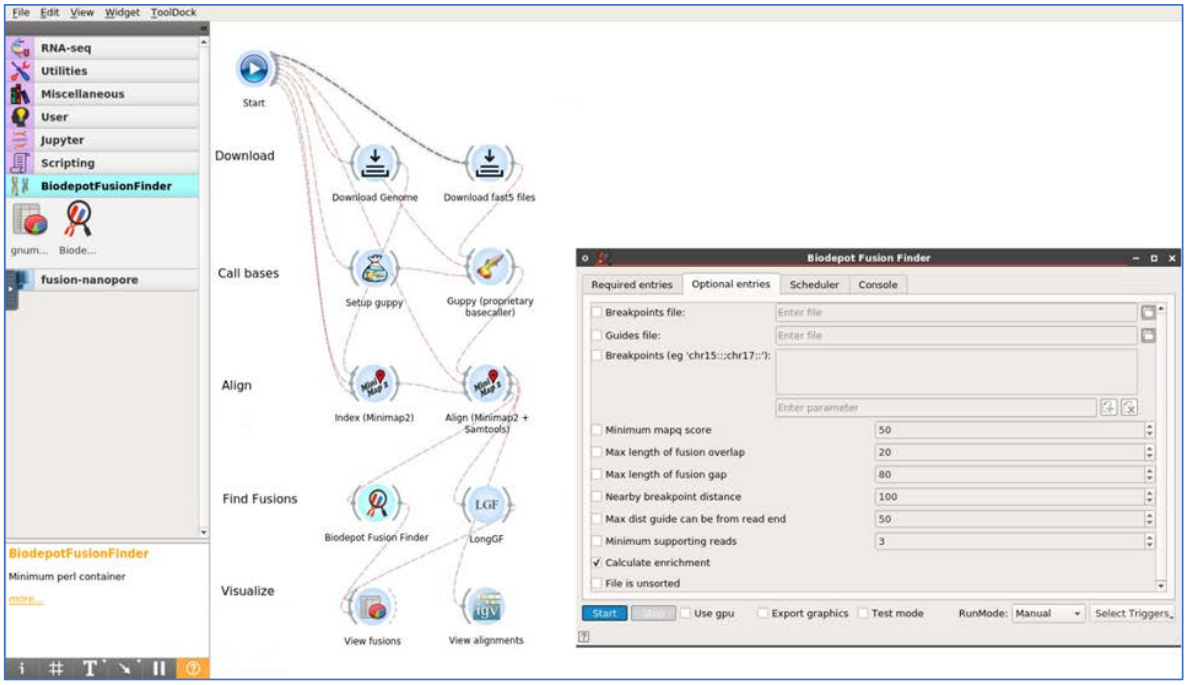
4.3.1 Deepvariant

The runtime benchmark comparison of Deepvariant CPU and GPU shows a noticeable speedup in the Inference step of the pipeline. This step benefits the most from GPU acceleration as it utilizes large Convolutional Neural Network (CNN) models for prediction of candidate variants. The runtime comparison between CPU and GPU is shown in Table 4.3 and Figure 4.2. Deepvariant GPU has a 2.8x speedup in the inference step and 1.5x speedup overall.

4.3.2 PEPPER-Margin-Deepvariant

The PEPPER-Margin-Deepvariant benchmarks better display the runtime differences between CPU and GPU-accelerated versions due to the larger nanopore dataset used. The runtime comparison between CPU and GPU is shown in Table 4.4 and Figure 4.3. Significant runtime differences can be seen in multiple stages of this pipeline over CPU and GPU-accelerated versions of the workflow. PEPPER-Margin-Deepvariant GPU achieves a 2.1x speedup in the PEPPER-Candidate stage, 2.8x speedup in Deepvariant's SNP stage, 2.5x speedup in Deepvariant's INDEL stage, and a 2.5x speedup overall. The containerized variant-calling workflow used in testing can be seen in Figure 4.4.

A: Bwb workflow with LongGF &BFF



B: Enrichment statistics computation output

	A	B	C	D	E	F	G	H	I	J
1	Breakpoint	Gap/Overlap	Count	Nearby-count	ReadThru	NearbyReadThru	Breakpoint enrichment	Breakpoint nearby enrichment	Fraction on target	Fraction on target nearby
2	chr9:133607156;chr22:23632742	2	1	29	30	31	30.19597	875.68318	0.03333	0.93548
3	chr9:133607152;chr22:23632739	4	1	29	29	31	30.19597	875.68318	0.03448	0.93548
4	chr9:133607147;chr22:23632742	5	21	29	28	31	634.11541	875.68318	0.75000	0.93548
5	chr9:133607164;chr22:23632742	5	1	29	31	31	30.19597	875.68318	0.03226	0.93548
6	chr9:133607168;chr22:23632748	5	1	29	30	31	30.19597	875.68318	0.03333	0.93548
7	chr9:133607158;chr22:23632736	8	1	29	31	31	30.19597	875.68318	0.03226	0.93548
8	chr9:133607147;chr22:23632715	16	1	29	31	31	30.19597	875.68318	0.03226	0.93548
9	chr9:133607177;chr22:23632748	17	1	29	30	31	30.19597	875.68318	0.03333	0.93548
10	chr9:133607153;chr22:23632728	20	1	29	31	31	30.19597	875.68318	0.03226	0.93548
11	Coverage	MaxBpFromEnd	GuideReads	Enrichment						
12	0.0331	50	212	6401.546						

C: IGV image confirming fusion

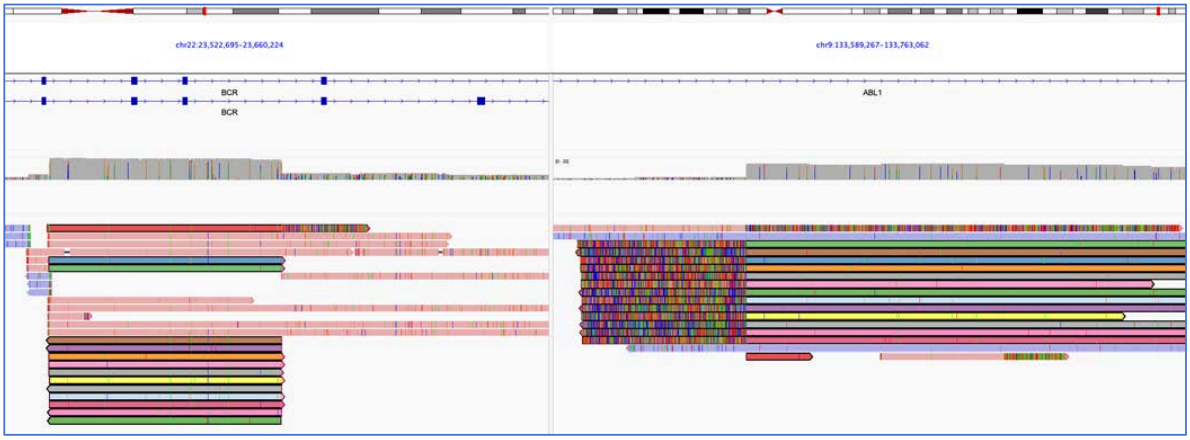


Figure 4.1: (Continued on the following page)

Figure 4.1: Screenshot of the nanopore fusion workflow and output. **Panel A**: Bwb workflow including our custom Biodepot Fusion Finder (BFF) and LongGF widgets. **Panel B**: shows enrichment statistics as fusion enrichment and on target enrichment. **Panel C**: BCR-ABL1 fusion viewed in Integrated Genomics Viewer (IGV).

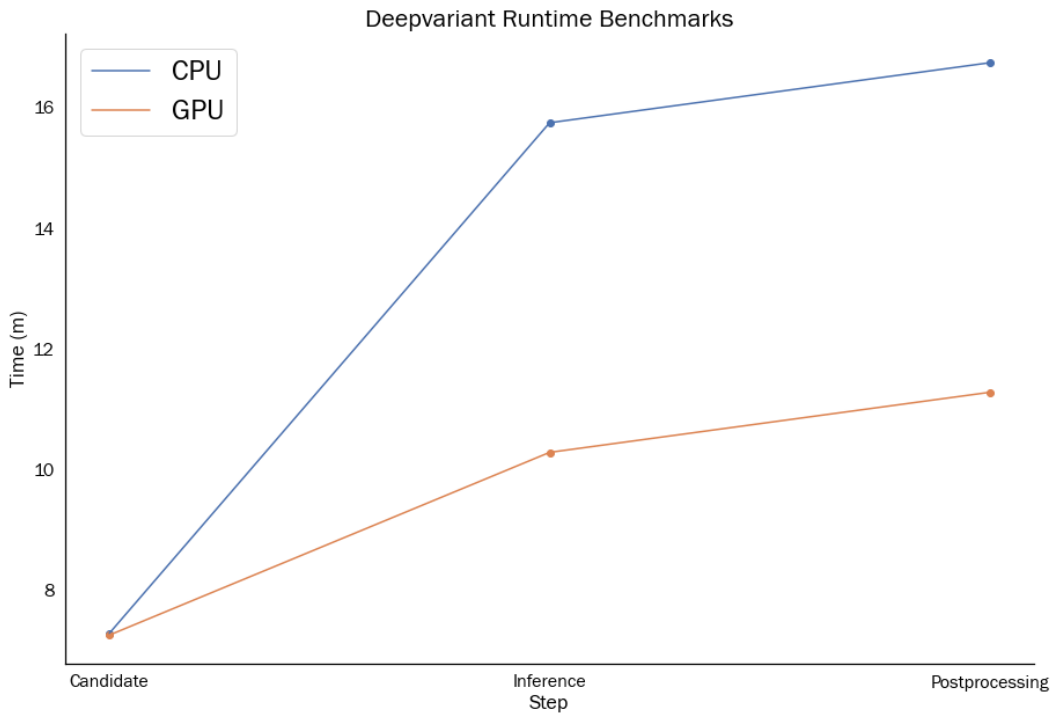


Figure 4.2: Deepvariant Runtime Benchmarks

Table 4.2: Comparison of LongGF and Biodepot-Fusion-Finder (BFF) fusions found.

	Sample Statistics			Fusions Found	
Sample Name	Gb of Data	Mean Coverage	On-target Enrichment	LongGF	BFF
K562	0.12	0.04	5830.00	0	29
KU812	0.36	0.11	2108.33	0	51
KCL22	0.69	0.21	2118.70	5	171
NB4	0.07	0.02	848.57	3	3
MV4;11	0.28	0.08	2698.93	57	71
ME1	0.91	0.28	1305.49	45	37
CML1	0.92	0.28	1563.91	0	142
CML2	0.04	0.01	2227.50	0	4
CML3	0.09	0.03	3006.67	0	0
CML4	0.93	0.28	557.10	0	14
CML5	0.04	0.03	1283.33	0	8
CML6	5.00	1.52	1539.12	0	0
AML1	5.47	1.66	535.12	28	38
AML2	1.59	0.48	649.62	0	0
APL1	0.17	0.05	1242.35	0	0
APL2	1.10	0.33	618.00	8	10
APL3	0.31	0.09	872.90	8	10
APL4	1.63	0.49	1785.64	0	0
APL5	0.22	0.07	2520.00	0	0
APL6	1.06	0.32	650.66	0	2

Table 4.3: Comparison of Deepvariant CPU and GPU.

	Step Runtimes (minutes)			
Variant Caller	Candidate	Inference	Postprocessing	Variantcall-Total
Deepvariant CPU	7.284	8.467	0.996	16.747
Deepvariant GPU	7.255	3.030	0.997	11.282

Table 4.4: Comparison of PEPPER-Margin-Deepvariant CPU and GPU.

	Step Runtimes (minutes)					
Variant Caller	PEPPER-Candidate	Margin-Phase	SNP	INDEL	Merge	Total
PEPPER CPU	29.118	1.932	86.915	23.214	0.572	143.072
PEPPER GPU	13.812	1.757	30.991	9.326	0.524	58.006

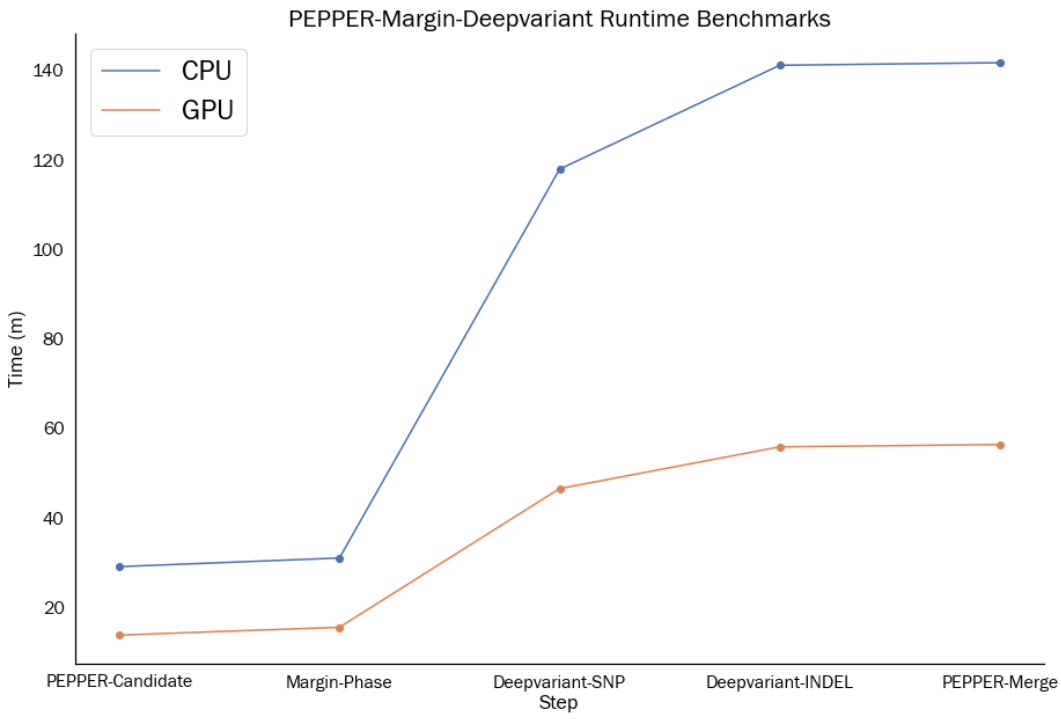


Figure 4.3: PEPPER-Margin-Deepvariant Runtime Benchmarks

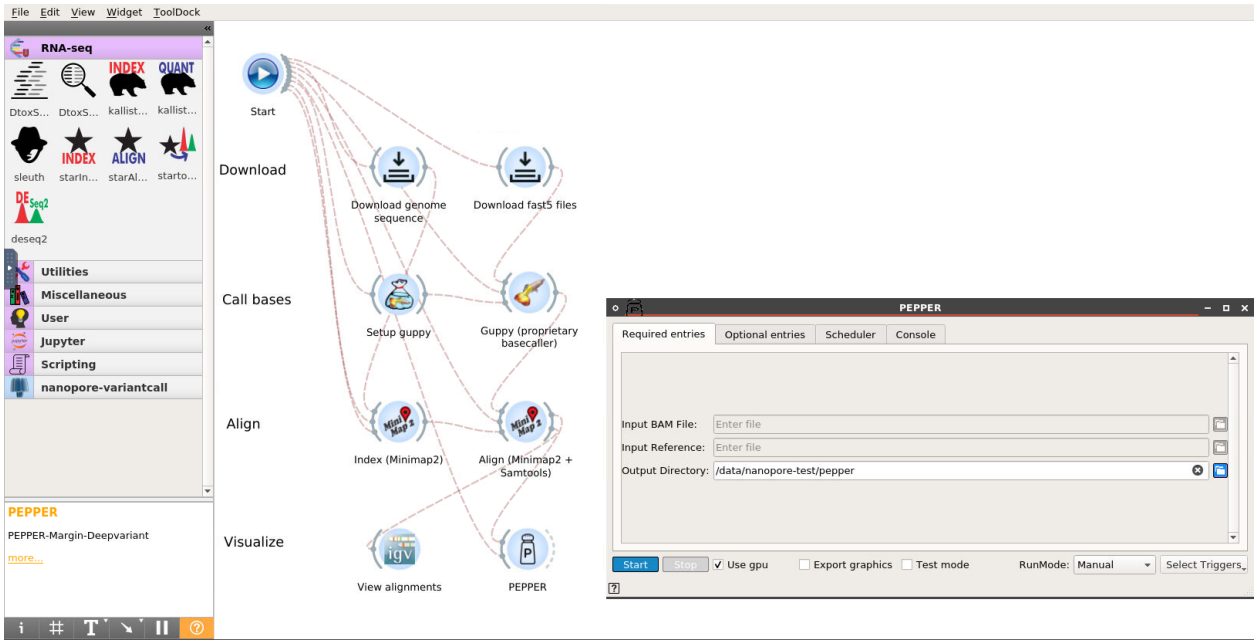


Figure 4.4: Screenshot of the variant-calling workflow containing PEPPER-Margin-Deepvariant.

Chapter 5

CONCLUSION

5.1 Statement

This thesis project focuses on computational methods and software tools tailored for long-read sequencing data. We have created three containerized, cloud-enabled, hardware-accelerated nanopore workflows that have been benchmarked with runtime and accuracy metrics. Our baseline workflow contains the proprietary basecaller Guppy in the basecalling step followed by Minimap2 for alignment and finally Integrative Genomics Viewer for visualization. We observed a 29x speedup in average runtime (in seconds) between the GPU and CPU versions of this workflow. Our fusion detection workflow focuses on rapid detection of myeloid neoplasm fusions and also supports calculation of a sample's enrichment. We show that our tool for detecting fusions – Biodepot-Fusion-Finder (BFF) – detects significantly more fusions than the state of the art fusion detection tool LongGF. Finally, our variant-calling workflow focuses on a complete, haplotype-aware variant calling pipeline starting with basecalling and ending with variant call format output. With runtime benchmarks, we observed a 2.5x speedup overall with GPU usage as compared to CPU.

BIBLIOGRAPHY

[1] Bonito. a pytorch basecaller for oxford nanopore reads [https://github.com/nanoporetech/bonito].

[2] Oxford nanopore technologies github: Guppy [https://github.com/nanoporetech].

[3] Shanika L. Amarasinghe, Shian Su, Xueyi Dong, Luke Zappia, Matthew E. Ritchie, and Quentin Gouil. Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, 21(1), 2020.

[4] Shanika L. Amarasinghe, Shian Su, Xueyi Dong, Luke Zappia, Matthew E. Ritchie, and Quentin Gouil. Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, 21(1), 2020.

[5] Olson ND;Wagner J;McDaniel J;Stephens SH;Westreich ST;Prasanna AG;Johanson E;Boja E;Maier EJ;Serang O;Jáspez D;Lorenzo-Salazar JM;Muñoz-Barrera A;Rubio-Rodríguez LA;Flores C;Kyriakidis K;Malousi A;Shafin K;Pesout T;Jain M;Paten B;Chang PC;Kolesnikov A;Nat. Precisionfda truth challenge v2: Calling variants from short and long reads in difficult-to-map regions.

[6] Daniel A. Arber, Attilio Orazi, Robert Hasserjian, Jürgen Thiele, Michael J. Borowitz, Michelle M. Le Beau, Clara D. Bloomfield, Mario Cazzola, and James W. Vardiman. The 2016 revision to the world health organization classification of myeloid neoplasms and acute leukemia. *Blood*, 127(20):2391–2405, 2016.

[7] Cumbo, Minervini, Orsini, Anelli, Zagaria, Minervini, Coccaro, Impera, Tota, Parciante, and et al. Nanopore targeted sequencing for rapid gene mutations detection in acute myeloid leukemia. *Genes*, 10(12):1026, 2019.

[8] John E. Gorzynski, Sneha D. Goenka, Kishwar Shafin, Tanner D. Jensen, Dianna G. Fisk, Megan E. Grove, Elizabeth Spiteri, Trevor Pesout, Jean Monlong, Gunjan Baid, Jonathan A. Bernstein, Scott Ceresnak, Pi-Chuan Chang, Jeffrey W. Christle, Henry Chubb, Karen P. Dalton, Kyla Dunn, Daniel R. Garalde, Joseph Guillory, Joshua W. Knowles, Alexey Kolesnikov, Michael Ma, Tia Moscarello, Maria Nattestad, Marco Perez, Maura R.Z. Ruzhnikov, Mehrzad Samadi, Ankit Setia, Chris Wright, Courtney J. Wusthoff, Katherine Xiong, Tong Zhu, Miten Jain, Fritz J. Sedlazeck, Andrew Carroll,

Benedict Paten, and Euan A. Ashley. Ultrarapid nanopore genome sequencing in a critical care setting. *New England Journal of Medicine*, 386(7):700–702, 2022.

[9] Karin Helmersen and Hege Vangstein Aamot. Dna extraction of microbial dna directly from infected tissue: An optimized protocol for use in nanopore sequencing. *Scientific Reports*, 10(1), 2020.

[10] Ling-Hong Hung, Jiaming Hu, Trevor Meiss, Alyssa Ingersoll, Wes Lloyd, Daniel Kristiyanto, Yuguang Xiong, Eric Sobie, and Ka Yee Yeung. Building containerized workflows using the biodepot-workflow-builder. *Cell Systems*, 9(5), 2019.

[11] T. Laver, J. Harrison, P.A. O'Neill, K. Moore, A. Farbos, K. Paszkiewicz, and D.J. Studholme. Assessing the performance of the oxford nanopore technologies minion. *Biomolecular Detection and Quantification*, 3:1–8, 2015.

[12] Mark Levis. Faculty opinions recommendation of diagnosis and management of acute myeloid leukemia in adults: Recommendations from an international expert panel, on behalf of the european leukemianet. *Faculty Opinions – Post-Publication Peer Review of the Biomedical Literature*, 2014.

[13] Heng Li. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, 2018.

[14] Qian Liu, Yu Hu, Andres Stucky, Li Fang, Jiang F. Zhong, and Kai Wang. Longgf: Computational algorithm and software tool for fast and accurate detection of gene fusions by long-read transcriptome sequencing. *BMC Genomics*, 21(S11), 2020.

[15] Sunali Mehta, Andrew Shelling, Anita Muthukaruppan, Annette Lasham, Cherie Blenkiron, George Laking, and Cristin Print. Predictive and prognostic molecular markers for cancer medicine. *Therapeutic Advances in Medical Oncology*, 2(2):125–148, 2010. PMID: 21789130.

[16] Margaret R. O'Donnell, Martin S. Tallman, Camille N. Abboud, Jessica K. Altman, Frederick R. Appelbaum, Daniel A. Arber, Eyal Attar, Uma Borate, Steven E. Coutre, Lloyd E. Damon, and et al. Acute myeloid leukemia, version 2.2013. *Journal of the National Comprehensive Cancer Network*, 11(9):1047–1055, 2013.

[17] Brittany C. Parker and Wei Zhang. Fusion genes in solid tumors: An emerging target for cancer diagnosis and treatment. *Chinese Journal of Cancer*, 32(11):594–603, 2013.

[18] Ryan Poplin, Pi-Chuan Chang, David Alexander, Scott Schwartz, Thomas Colthurst, Alexander Ku, Dan Newburger, Jojo Dijamco, Nam Nguyen, Pegah T Afshar, and

et al. A universal snp and small-indel variant caller using deep neural networks. *Nature Biotechnology*, 36(10):983–987, 2018.

[19] PrecisionFDA. Truth challenge v2: Calling variants from short and long reads in difficult-to-map regions, 2020.

[20] Jerald Radich, Cecilia Yeung, and David Wu. New approaches to molecular monitoring in cml (and other diseases). *Blood*, 134(19):1578–1584, 2019.

[21] Shishir Reddy, Ling-Hong Hung, Olga Sala-Torra, Jerald P. Radich, Cecilia CS Yeung, and Ka Yee Yeung. A graphical, interactive and gpu-enabled workflow to process long-read sequencing data. *BMC Genomics*, 22(1), 2021.

[22] Peter Robinson and Tomasz Zemo jtel. Integrative genomics viewer (igv): Visualizing alignments and variants. *Computational Exome and Genome Analysis*, page 233–245, 2017.

[23] Olga Sala-Torra, Shishir Reddy, Ling-Hong Hung, Lan Beppu, David Wu, Jerald Radich, Ka Yee Yeung, and Cecilia CS Yeung. Rapid detection of myeloid neoplasm fusions using single molecule long-read sequencing. 2022.

[24] Fritz J. Sedlazeck, Hayan Lee, Charlotte A. Darby, and Michael C. Schatz. Piercing the dark matter: Bioinformatics of long-range sequencing and mapping. *Nature Reviews Genetics*, 19(6):329–346, 2018.

[25] Kishwar Shafin, Trevor Pesout, Pi-Chuan Chang, Maria Nattestad, Alexey Kolesnikov, Sidharth Goel, Gunjan Baid, Mikhail Kolmogorov, Jordan M. Eizenga, Karen H. Miga, and et al. Haplotype-aware variant calling with pepper-margin-deepvariant enables high accuracy in nanopore long-reads. *Nature Methods*, 18(11):1322–1332, 2021.

[26] Maria S. Tretiakova, Wenjing Wang, Yu Wu, Scott S. Tykodi, Lawrence True, and Yajuan J. Liu. Gene fusion analysis in renal cell carcinoma by fusionplex rna-sequencing and correlations of molecular findings with clinicopathological features. *Genes, Chromosomes and Cancer*, 59(1):40–49, 2019.

[27] Paul A. VanderLaan, Yigu Chen, Marcello DiStasio, Deepa Rangachari, Daniel B. Costa, and Yael K. Heher. Molecular testing turnaround time in non–small-cell lung cancer: Monitoring a moving target. *Clinical Lung Cancer*, 19(5), 2018.

[28] Ryan R. Wick, Louise M. Judd, and Kathryn E. Holt. Performance of neural network basecalling tools for oxford nanopore sequencing. *Genome Biology*, 20(1), 2019.

[29] Xin Xiao, Cassandra C. Garbutt, Francis Hornicek, Zheng Guo, and Zhenfeng Duan. Advances in chromosomal translocations and fusion genes in sarcomas and potential therapeutic applications. *Cancer Treatment Reviews*, 63:61–70, 2018.

[30] Marjan Yaghmaie and Cecilia CS Yeung. Molecular mechanisms of resistance to tyrosine kinase inhibitors. *Current Hematologic Malignancy Reports*, 14(5):395–404, 2019.

[31] Cecilia C. Yeung and Jerald Radich. Predicting chemotherapy resistance in aml. *Current Hematologic Malignancy Reports*, 12(6):530–536, 2017.