

Evaluation of mapping and germline variant calling pipelines on Australian high-performance computing facilities

Georgina Samaha, Tracy Chew, Cali Willet, Sarah Beecroft, Brian Davis, Rosemarie Sadsad

Executive summary

Whole genome sequencing (WGS) is the largest and most commonly stored data type across 27 organisations surveyed in Australia (Australian Genomics Health Alliance, 2020). Of these organisations, 82% use command-line interface (CLI) platforms. As sequencing becomes more affordable, an increasing number of life-scientists are using these technologies at scale. Australian life-science researchers seek best practice pipelines that are accurate, highly accessible, well documented and are accompanied with user support or training. Australian researchers experience challenges with deploying these pipelines on local high performance computing (HPC) infrastructures at scale. This is largely because many best practice tools and workflows were not developed for large scale use, or for HPC. Recently, scalable pipelines to process and analyse WGS data have been developed and made publicly available (Chew et al. 2021; NVIDIA 2020; Willet et al. 2021). Scalability of these pipelines on HPC infrastructure is achieved by re-engineering best practice pipelines to efficiently utilise compute hardware and by replacing recommended tools with tools that are more computationally performant. This includes the use of computational strategies such as code parallelisation and use of specialised hardware such as graphical processing units (GPUs). Proper technical evaluation is required to determine whether their biological accuracy is maintained alongside improvements in computational efficiency.

Here we report the biological accuracy and technical performance of two scalable WGS pipelines that perform short read mapping to a reference genome assembly (herein referred to as ‘mapping’) and germline short variant discovery of single nucleotide variants (SNVs) and insertions and deletions (indels). Both workflows are implementations of the BROAD Institute’s best practices pipelines: “Data pre-processing for variant discovery” and “Germline short variant calling (SNVs + indels)” (De Pisto et al. 2011; McKenna et al. 2010) and were run using their respective default or recommended settings. These workflows are widely adopted in the community (Zhao et al. 2020). We used the gold standard Platinum Genomes datasets (Eberle et al. 2016) to report metrics to evaluate: 1. NVIDIA’s Clara Parabricks GPU-enabled Pipelines (NVIDIA 2020) and 2. The Sydney Informatics Hub’s (SIH) Scalable multi CPU node pipelines (Chew et al. 2021; Willet et al. 2021), both deployed on the National Computational Infrastructure (NCI) HPC ‘Gadi’ [<https://nci.org.au/our-systems/hpc-systems>]. This report is primarily focused on comparing the biological accuracy of each pipeline, run using default settings. Technical benchmarks and scalability estimations are presented specifically in the context of NCI’s Gadi HPC. Scalability testing was not performed as we were limited to six

samples and both Parabricks and SIH pipelines have previously exhibited a capacity to scale (AWS Editorial Team et al. 2021; Franke & Crowgey, 2020; Hayes et al. 2021; Lott et al. 2022; Satgunaseelan et al. 2021). In addition to biological accuracy, we evaluate user friendliness, theoretical ability to scale on NCI's Gadi HPC, flexibility to modify, and the user support model of each pipeline. This report is intended to guide users with their choice of workflow based on the resources made available by the Australian Biocommons and at NCI Gadi.

A comparison of overall performance by Parabricks and SIH pipelines is presented in Table 1. The Parabricks pipeline captured more true positive (2,010,754) and false positive SNV and indel variants (70,000) than the SIH pipeline (1,963,111 and 50,535, respectively). Parabricks returned substantially more SNVs (7,012,334) and indels (1,419,947) compared with the SIH pipeline (6,074,524 SNVs, 1,475,260 indels). The SIH pipeline produced more precise SNV callsets (0.987 ± 0.01), compared with the Parabricks pipeline (0.972 ± 0.01). Recall of SNVs was higher for Parabricks (0.984 ± 0.01), compared with SIH (0.958 ± 0.01), and comparable for indels for both Parabricks (0.958 ± 0.01) and SIH pipelines (0.955 ± 0.01). F1 scores of indel calling were higher for the Parabricks pipeline (0.979 ± 0.01) compared with the SIH pipeline (0.942 ± 0.01). Parabricks and SIH pipelines used different tools for joint genotyping of variants, which is an important step in reducing false positive signals. The GLnexus joint genotyping tool (Lin et al. 2018) used by Parabricks has previously been shown to improve variant calling accuracy in WGS compared with the GATK tools used by the SIH pipeline (Yun et al. 2020). Despite this, the difference in biological accuracy between the pipelines was minimal and users may consider cost (as Parabricks is commercially licensed), capacity to scale, flexibility to adjust the pipeline, and user experience when choosing a germline variant calling implementation at NCI Gadi.

The overall runtime, service unit (SU) cost, and efficiency of mapping and germline variant calling pipelines is lower for the GPU-based Parabricks implementation (623.81 SUs/sample), compared with the SIH CPU-based pipeline (899.03 SUs/sample). The use of more efficient hardware like GPUs can reduce the carbon footprint of whole genome mapping and variant calling pipelines, without compromising on performance. At NCI Gadi, a single Parabricks license can process a maximum of 40 samples in a single run for both mapping and germline variant calling processes, given a maximum queue limit request of 960 CPUs and five hours. The SIH pipeline is capable of processing 506 and 144 samples in a single run for mapping and germline variant calling, respectively, using a total of 60,720 and 72,864 CPUs. NCI Gadi has 640 GPUs and 155,000 CPU cores. As it is commercially licensed software, Parabricks has additional limitations placed on its capacity for parallelism and scalability across cores available at NCI Gadi, compared with the multi-node capable SIH pipeline. While GPUs are designed for highly parallel computations, the use of interval chunking and scatter-gather approaches by the SIH pipeline offered improved throughput, runtime, and efficiency of mapping (96.74 mins for SIH compared with 177.66 mins for Parabricks) and variant calling (37.17 mins for SIH compared with 89.85 mins for Parabricks) steps. However, this is only made possible through the availability of a large number of CPUs at NCI Gadi.

Parabricks provides a simplified commercially licensed toolkit for users and offers the choice of running the complete pipeline with one command. The caveat to this is that the users' ability to

customise commands and parameters for their needs is limited at times. For example, Parabricks does not allow users to perform alternate contig aware (ALT-aware) mapping which is used to improve variant calling across highly variable regions of the genome. The SIH pipeline consists of a series of scripts to be run manually by the user. These scripts can be customised according to the user's needs and have been optimised for both human and non-human datasets but assume a moderate level of CLI proficiency. User preference for the mapping and variant calling pipelines offered by Parabricks or SIH at NCI Gadi will depend on their CLI experience, need to customise the workflow to directly address their research questions, and the size of their dataset.

Table 1: Overall performance comparison of mapping and germline variant calling pipelines run on six Platinum Genomes samples on National Compute Infrastructure's (NCI) Gadi supercomputer.

	NVIDIA Clara Parabricks	Sydney Informatics Hub Pipelines
Version	3.5	2.0
Infrastructure	NCI Gadi	NCI Gadi
Requirements or dependencies	Software license CUDA \geq 9.0 2 GPUs, 24 CPU, 100 GB RAM Python3 NVIDIA-Docker (or Singularity \geq 2.6.1)	Nci.parallel OpenMPI v4.1 BWA-MEM v0.7.17 BWAKit v0.7.11 Fastp v0.20.0 FastQC v0.11.7 GATK v4.1.8.1 Java jdk 1.8.0_60 K8 v0.2.5 Sambamba v0.7.1 Samblaster v0.1.24 Samtools v1.10 Seqtk v1.3-r113-dirty
Portability	Moderately portable	Not portable
Flexibility to adjust parameters to suit project requirements	Somewhat flexible	Highly flexible
Total CPUs used	456 (20 GPUs)	960
End to end walltime for 6 samples (mins) (FASTQ to VCF)	1,567.71*	531.94
Service Units Per Sample	623.81	899.03
Scalability limits	40 samples/batch [†]	Mapping: 506 samples/batch [†] Germline variant calling: 144 samples/batch [†]
SNV calling accuracy (F1- score)	0.978 \pm 0.011	0.972 \pm 0.010
Indel calling accuracy (F1- score)	0.941 \pm 0.014	0.930 \pm 0.015
User support	Available	Limited

User experience	Few commands Easily customised Commercially licensed	User input required at each step of the pipeline Easily customised for proficient CLI users
Unique features	Can be modularised Can run germline pipeline in single command	Checkpointing ALT-aware calling Reports variant calling metrics Outputs split and discordant reads for structural variant calling

* Parabricks was run using 2 GPUs, 2 trial licenses and 1 GPU node. See 'Pipeline evaluation methods' section in supplementary information for more details.

[†] a 'batch' is defined as the number of samples that can be processed in one run of the pipeline, considering resource request limits imposed by NCI for each queue. Maximum resource request is 20,736 CPUs and 5 hours walltime for the normal queue and 960 CPUs and 5 hours for the gpuvolta queue.

Scientific context of mapping and variant calling pipelines

Whole genome sequencing is the process of determining an organism's entire DNA sequence. WGS is used to discover genetic variants including SNVs and indels that are changes in the DNA sequence. In humans, these variations determine an individual's unique genomic profile and mostly have no observable effect. However, some are potentially harmful to normal biological function and result in heritable conditions including cancer, haemophilia, and albinism, and others contribute to non-harmful differences in phenotypic traits. Identifying both harmful and benign variant sites is a critical step in genetic research, upon which all downstream analyses and interpretation depend.

Most organisms have genomes that are too large to be sequenced end-to-end in a single step. WGS involves first deconstructing an individual's genome into smaller pieces, which are then sequenced computationally in a multi-step process (Figure 1). Before DNA can be sequenced, it is chemically broken down into small fragments that can be processed by sequencing machines in a laboratory. These fragments are then read by a sequencing machine and the order of nucleotide bases (A, T, C, and G) is determined for each one. These sequenced fragments are referred to as 'reads'. WGS reads are then reassembled computationally to deduce the original genome sequence. This involves mapping of reads to a reference genome, which is a species-specific standard that serves as a 'framework' for organising the sequence reads into a genomic sequence. Reads are mapped to the reference and sorted by their genomic location to build a continuous sequence. This mapped sequence is further processed to remove systematic technical errors introduced during the sequencing process (marking of duplicate reads and base quality score recalibration). Variant sites are then called from an individual's sequence, relative to the reference (sample-level variant calling). WGS studies typically require multiple individuals to reliably identify variant sites associated with a trait of interest. As such, individuals in a study cohort are processed together in a series of steps (cohort-level joint variant calling and

post-processing). This final step produces consistent genotype information across the cohort that makes it easier to accurately identify variant sites.

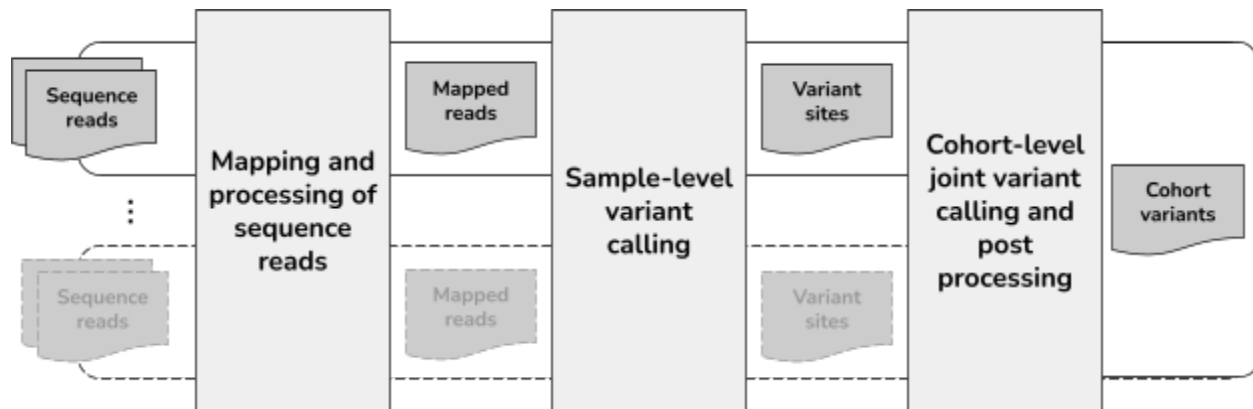


Figure 1: Overview of the stages of the BROAD Institute's Best Practice recommendations for mapping and germline variant calling of whole genome sequence data for individual samples.

Each of these stages is composed of a series of processes that are performed by different software, pieced together into a workflow. As WGS has become an integral method for life science and clinical research, multiple tools have been developed to process WGS data and identify germline variants. Currently, GATK Best Practice workflows developed by the BROAD Institute (De Pisto et al. 2011) are widely applied and popular among the genomics community. These recommendations comprise analysis phases that are mentioned above, including data pre-processing (mapping and processing of sequence reads) in which raw sequence reads in FASTQ format are mapped to a reference genome, and mapped sequence reads are processed in the format of BAM files. Sample-level variant calling is performed on these BAM files to produce variant calls in the genome VCF (gVCF) format, which are then processed as a cohort to produce a VCF file that contains genotype information for all variant sites across all individuals in the cohort.

Both Parabricks and SIH mapping and germline variant calling pipelines evaluated in this report followed The BROAD Institute's Best Practice recommendations, with slight differences in tool utilisation and parameters accounting for differences in final results. Please see 'Pipeline details' within the supplementary information for further details on individual steps run by each pipeline.

Biological accuracy

The biological accuracy of mapping and variant calling pipelines is commonly determined by their ability to differentiate true variants from sequencing artifacts that lead to false positive calls. Two metrics were used here to measure the biological accuracy of Parabricks and SIH pipelines. Recall metrics presented below capture the absence of false negatives, and precision metrics capture the absence of false positives in each pipeline's resulting variant set.

While best practice variant calling tools and methods are designed to achieve high recall and precision scores, different tools following the same workflow can achieve different results due to a number of reasons. The underlying statistical models used by various tools rely on different assumptions, the direct impacts of which can be difficult to evaluate experimentally. Further, parameters and thresholds can be adjusted for specific tools, further influencing the results (Pabinger et al. 2014). Finally, in order to assess the biological accuracy of these tools, we rely on standardised benchmarking datasets like Platinum Genomes, which are biased toward high-confidence regions of the genome in order to more reliably profile precision and recall. Each of these factors influenced the results presented here and limit our ability to accurately capture the performance of the pipelines evaluated here across the whole genome. Various implementations of the BROAD Institute's best practice GATK pipelines have previously been shown to achieve a high degree of accuracy in distinguishing true variants from false signals. Using the same gold standard benchmarking dataset used here, best practice GATK variant calling pipelines have achieved 0.954 recall, 0.998 precision, and 0.975 F1-score for SNVs and 0.939 recall, 0.987 precision, and 0.962 F1-score for indels (Zhao et al. 2020).

Both Parabricks and SIH mapping and germline variant calling pipelines evaluated in this report followed The BROAD Institute's Best Practice recommendation. The biological validity of the Parabricks and SIH pipelines was evaluated by comparing the concordance of SNVs and indels produced by each pipeline to the Platinum Genomes variant truth sets (See 'Platinum Genomes dataset' in Supplementary information). The biological accuracy of each pipeline was consistent across all samples. The Parabricks callset contained substantially more variants compared with the SIH callset (Table 2). This is likely attributed to the difference in joint variant calling tools used by the Parabricks (GLNexus) and SIH pipelines (GenoimcsDBImport) and filtering of variants based on variant quality score recalibration (VQSR) applied by the SIH pipeline's joint variant calling step. Across high confidence regions of the Hg38 genome assembly, the Parabricks pipeline consistently captured a greater proportion of true positive calls in the Platinum Genomes dataset, across both SNVs and indels, compared with the SIH pipeline (Figure 2a), however this increased sensitivity came at a cost of slightly reduced specificity compared to the SIH pipeline. Over 30% of the variants shared by the Parabricks and SIH callsets were not present in the Platinum Genomes truth set, indicating inadequate variant filtering was performed by both Parabricks and SIH pipelines. Variant filtering requirements differ between datasets and research questions, as such it is often performed separately and to varying degrees by end-users. Due to the highly flexible nature of the SIH pipeline, users are able to tweak VQSR parameters to achieve a level of sensitivity and specificity suitable for their project

(<https://gatk.broadinstitute.org/hc/en-us/articles/360035531612-Variant-Quality-Score-Recalibration-VQSR->). It is also important to note that the Platinum Genomes truth callsets were subjected to stringent quality filtering, as such it is possible that a proportion of these 'false' variants are true positives, that were excluded from the final truth callsets.

Table 2: Summary of final callsets produced by Parabricks and Sydney Informatics Hub (SIH) pipelines. Recall and precision are presented as average (\pm standard deviation) across six Platinum Genomes samples.

Pipeline	Total variants	Total SNVs	Total indels	Recall		Precision	
				SNV	Indel	SNV	Indel
Parabricks	8,371,855	7,012,334	1,419,947	0.984 \pm 0.01	0.958 \pm 0.01	0.972 \pm 0.01	0.924 \pm 0.01
SIH	7,517,673	6,074,524	1,475,260	0.958 \pm 0.01	0.955 \pm 0.01	0.987 \pm 0.01	0.905 \pm 0.02

The Parabricks pipeline demonstrated improved recall or ‘absence of false negatives’ for both SNVs and indels, compared with the SIH pipeline (Table 2; Figure 2b). The Parabricks pipeline also demonstrated improved precision or ‘absence of false positives’ in detecting indels, but not SNVs, compared with the SIH pipeline. The differences in accuracy of the Parabricks and SIH pipelines, while minimal, can be attributed to the difference in tools used for joint variant calling which is an essential step in reducing false positive signals. This is consistent with other studies that have shown the GLnexus joint variant calling tool used in the Parabricks pipeline to be more accurate across high-confidence regions of the genome than those used in the SIH pipeline in WGS datasets (Yun et al. 2020; Zhao et al. 2020).

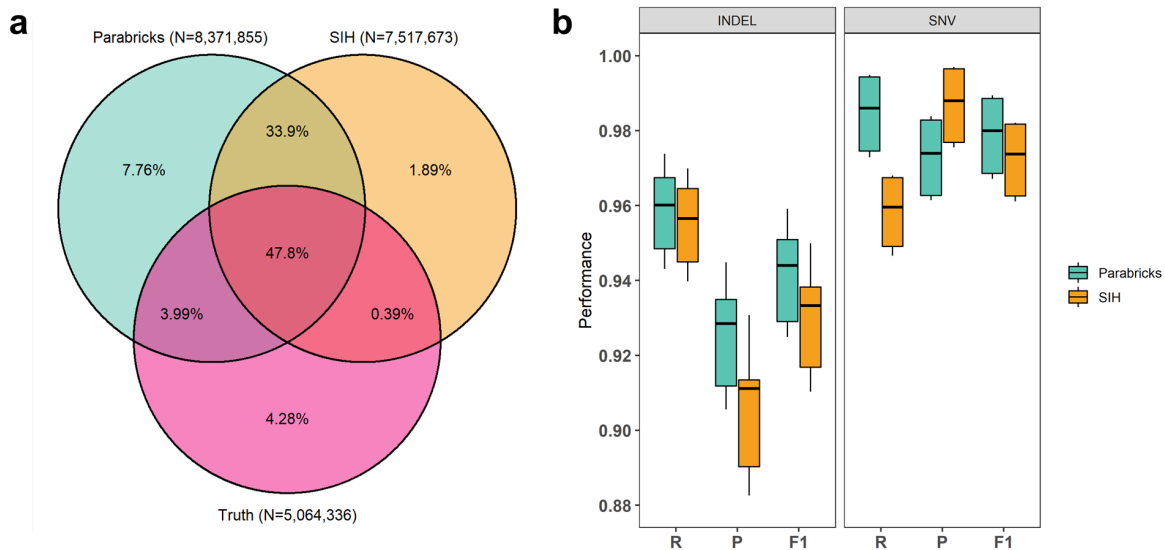


Figure 2: Truth variant representation across high confidence regions of the Hg38 reference genome. **a.** Venn diagram of variants shared by Parabricks SIH, and truth set for NA12878. **b.** Comparison of Precision (P), recall (R), F1-score (F1) for SNVs and indels across all samples. F1-score is the harmonic mean of precision and recall.

Unlike high-confidence regions of the genome, complex regions include those that are repeat rich or highly variable between individuals and pose additional challenges to NGS data analysis. These regions include clinically relevant parts of the genome such as the Human Leukocyte Antigen (HLA) complex. The HLA locus plays a vital role in adaptive and innate immunity, infectious and autoimmune diseases, and drug response. It is challenging to characterise the HLA locus with short read sequencing and reference-based variant calling methods due to its high degree of variability between individuals (Dos Santos et al. 2015). The GRCh38/hg38 genome assembly includes alternate (ALT) contigs that capture highly diverse regions including the HLA complex, enabling accurate mapping and genotyping of this region. The SIH pipeline's implementation of ALT-aware mapping offers researchers the ability to more accurately characterise this complex region, among other regions on the ALT contigs. Further, ALT-aware mapping can eliminate false positive signals across these regions in the primary assembly. Parabricks does not offer ALT-aware mapping and as such, limits researcher's ability to capture the variant profile among this region (Table 3). Researchers interested in regions covered by the ALT contigs will benefit from the use of the SIH pipeline, compared with the Parabricks pipeline. Given the biological accuracy of Parabricks and SIH pipelines was comparable, user preference for either one would depend on the genomic regions of interest and preference of joint genotyping tools.

Table 3: Representation of variants across the highly variable HLA complex is affected by use of ALT-aware mapping by the SIH pipeline, compared with the Parabricks pipeline. Truth callsets are limited to high confidence regions and do not include ALT contigs.

Callset	Variants called at primary-HLA locus	Variants called at ALT-HLA	Primary-HLA gene coding variants
Parabricks	50,142	0	540
SIH	38,778	2,630	284
Truth	41,426	NA	1,594

Technical performance

Both pipelines were run according to the resource recommendations of their respective developers. The resources available to the Parabricks pipeline were limited to one GPU node (four GPUs, 48 CPUs). Parabricks commands were run using the recommended minimum of two GPUs, 24 CPUs, as supported by one trial license. The SIH pipeline had access to 3,074 nodes, each containing 48 CPU cores. The relative technical performance of each pipeline was evaluated by comparing the wall time, memory usage, compute efficiency and service unit (SU) consumption of individual processes comprising each pipeline. Total SUs consumed by Parabricks (3742.87 SU) was 69.4% of the SIH-Gadi pipeline (5394.2 SU). End to end walltime

of all processes was significantly lower for the Parabricks pipeline, which ran 1.8 times faster than the SIH pipeline and consumed fewer resources (Figure 3). With access to only one license, we were restricted to running Parabricks with a single GPU node. The cumulative walltime for the Parabricks pipeline when run serially for the six Platinum Genome samples was 1567.71 minutes, 2.95 times longer than the SIH pipeline, which is designed to batch process samples in parallel.

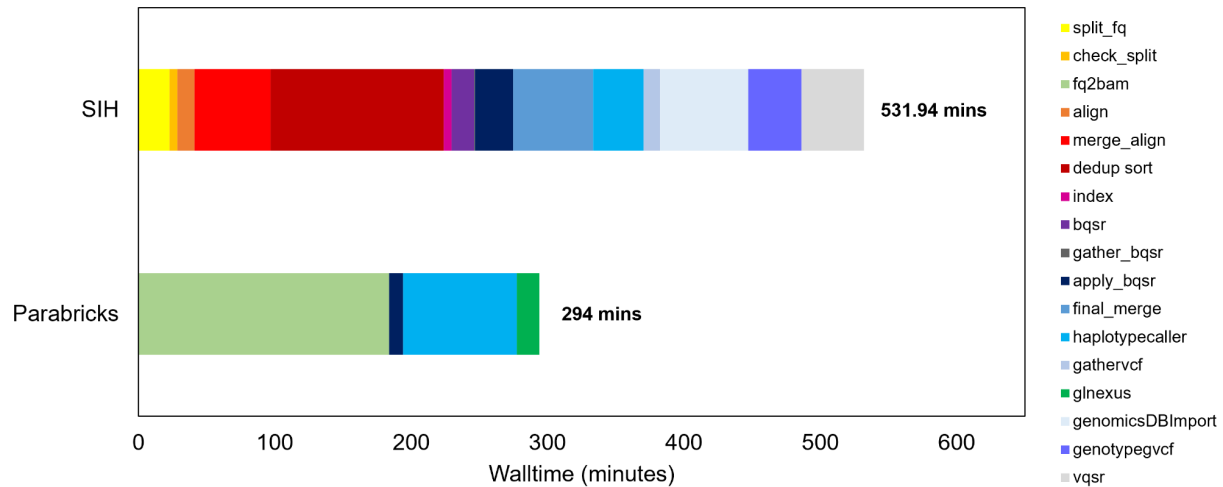


Figure 3: Cumulative and proportional walltimes of all processes comprising the SIH and Parabricks mapping and germline variant calling pipelines.

On NCI Gadi, Parabricks requires users to run all processes on the gpuvolta queue. The SIH pipeline has been optimised to run different processes on the normal and normalbw queues, and can also utilise the express queues when results are deemed sufficiently urgent to justify the additional SU cost. The most computationally complex processes of mapping and variant calling consumed the greatest proportion of SUs for both Parabricks and SIH pipelines (Table 4). However, the SIH pipeline achieved faster runtimes for both these processes, while consuming a greater number of SUs. Parabricks fq2bam command comprises a number of processes run by various steps in the SIH pipeline including: read alignment and sorting, marking of duplicate reads, and generation of BQSR tables. Of the 183.7 minutes taken by Parabricks' fq2bam command, 140.4 minutes were spent mapping reads to the reference assembly. Compared with the steps of splitfq, check_split, align, merge_align run by the SIH pipeline which took 96.74 minutes, Parabricks is slower than SIH to perform read alignment.

To ameliorate issues with the resource demands of the mapping process, the SIH pipeline uses a 'scatter-gather' approach to divide fastq files into even sized chunks, align those chunks in parallel and merge the aligned reads together. By doing this, the SIH pipeline (96.74 mins) achieves a faster turnaround time for alignments at a higher CPU efficiency than Parabricks' GPU-accelerated implementation of BWA-MEM (177.66 mins). Similarly, genomic interval chunking coupled with scatter-gather operations by the SIH haplotypcaller process improved walltime of variant calling with GATK's HaplotypeCaller, compared with the GPU-optimised

Parabricks haplotypcaller process. Overall, the Parabricks pipeline produced cohort-level VCFs in a shorter timeframe and lower SU cost, compared with the SIH pipeline. The differences presented here are a reflection of the additional processing steps of VQSR, implemented by the SIH pipeline, the choice of tools used by each pipeline and differences in GPU and CPU processing capacity.

Table 4: Resource requirements of all jobs comprising the SIH and Parabricks pipelines. All Parabricks jobs were run on the gpuvolta queue which has a charge rate of 3 SU/hour. SIH jobs were run on normal and normalbw queues with charge rates of 2 SU/hour and 1.25SU/hour respectively. Memory and service units (SU) are reported as the total consumed for running six samples, walltime is reported as maximal value across all samples.

Pipeline	Process*	NCPU used	Job queue	RAM (GB) used	Walltime (mins)	CPU efficiency	SU	% total SU
Parabricks	Fq2bam	24*	gpuvolta	1882.73	183.77	0.61	2380.37	63.60
	applybqsr	24*	gpuvolta	867.9	10.2	0.86	132.03	3.53
	haplotypcaller	24*	gpuvolta	741.25	89.85	0.99	1190.28	31.80
	glnexus	24*	gpuvolta	168.34	16.83	0.28	40.19	1.07
SIH	split_fq	96	normal	198.2	22.8	0.80	72.96	1.35
	check_split	96	normal	189.87	5.6	0.72	17.92	0.33
	align	5760	normal	18470	12.82	0.79	2460.8	45.62
	merge_align	168	normalbw	687.34	55.52	0.37	194.31	3.60
	dedup sort	168	normalbw	1300	126.88	0.21	451.14	8.36
	index	144	normal	281.75	5.82	0.34	27.92	0.52
	bqsr	384	normal	1370	16.47	0.42	210.77	3.91
	gather_bqsr	6	normal	22.44	0.82	0.83	0.16	0.00
	apply_bqsr	720	normal	1080	27.97	0.41	671.2	12.44
	final_merge	6	normal	113.64	58.55	0.86	26.02	0.48
	haplotypcaller	864	normal	1790	37.17	0.88	1070	19.84
	gathervcf	1	normalbw	16.43	11.85	0.71	9	0.17
	genomicsDBImport	28	normalbw	85.18	64.7	0.97	38	0.70
	genotypegvcf	96	normal	181.36	39.25	0.59	126	2.34
	vqsr	1	normal	28.73	45.72	0.99	18	0.33

* See 'Pipeline details' in Supplementary Information for details of commands run for each process.

User experience

As life scientists increasingly incorporate NGS sequencing into their research practices, it is essential that bioinformatics pipelines meet an expanding spectrum of need, while reducing the

labour, time management and maintenance by users. Bioinformatics end users are often wet lab biologists or clinicians with minimal computational experience, especially in the HPC setting. From the end user perspective, bioinformatics pipelines should be accessible, reliable, easy to use, well documented, and provide a foundation for reproducibility of published research findings. On top of this, challenges of working with HPC architectures, including software installation, job scheduling, and resource management, make running these pipelines more difficult for end users (Perez-Wohlfeil et al. 2018). Parabricks and SIH pipelines both offer reproducible workflows that have implemented versioning of both the pipeline and composite tools, simplified installation, and are comprehensively documented. The Parabricks pipeline comprised four separate processes, executed as 19 jobs for all six samples. The SIH pipeline comprised 15 separate processes, run as 15 jobs for all six samples. Given the differences in their implementation and user interface, user preference for one pipeline over another will depend on a number of factors including the need to scale, HPC and CLI experience, the user's preferences for command customisation, and the research questions being asked.

WGS pipelines typically consist of multiple tools with varying resource needs, pieced together into a workflow. As such, installing a pipeline's requisite tools and their dependencies is a complicated process for users with limited computational experience and understanding. Both Parabricks and SIH pipelines have addressed this through simplified installation processes. Parabricks installation is container-based and offers portability across infrastructures that is not achievable with the SIH pipeline. The SIH pipeline has been developed and optimised specifically for NCI Gadi, and all tools required by the pipeline are pre-installed in NCI Gadi's shared apps directory. Further, all scripts and reference files for the SIH pipeline can be downloaded from the SIH-Bioinformatics GitHub repository. Running these pipelines at NCI Gadi requires the use of the PBS Pro job scheduler and OpenMPI (Graham et al. 2006) to distribute tasks across the compute infrastructure. This can be challenging for users that lack HPC experience. The SIH pipeline scripts provide all relevant PBS job directives and preconfigured OpenMPI commands to distribute parallel jobs across nodes. Users of the SIH pipeline are required to perform minimal editing of some PBS directives to run their scripts, for example increasing the number of nodes requested depending on the number of samples to be processed in parallel. While Parabricks offers a simplified toolkit that reduces the number of commands to generate cohort variant callsets, users must still be familiar enough with high performance computing to interact with the job scheduler.

The lowering cost of WGS has rendered large-scale projects possible, leading to massive datasets that are used for a range of research applications across multiple biomedical disciplines. Life science researchers increasingly turn to national HPC facilities to efficiently process these growing volumes of data. As such, the scalability of bioinformatics pipelines optimised for these infrastructures is essential to their utility. The scalability of variant calling pipelines refers to their capacity to process multiple samples in parallel. The scalability of Parabricks and the SIH pipelines is reported here in the context of the 640 GPU and 155,000 CPU cores available at NCI Gadi. Parabricks is limited in its scalability relative to the SIH pipelines because of the number of available GPU nodes compared with CPU nodes and licensing requirements (figure 4). Given resource request limits imposed by NCI (20 GPU

nodes, 5 hours), Parabricks is capable of processing 40 samples in parallel at NCI Gadi. However, this would require access to 40 Parabricks licenses. A single Parabricks license is capable of processing a maximum of five samples per day at NCI Gadi. With 20 CPU nodes, the SIH Fastq-to-BAM pipeline is capable of processing three times as many samples (figure 4). The implementation of a ‘scatter-gather’ approach to resource distribution for data intensive processes like read mapping in the SIH Fastq-to-BAM pipeline is responsible for this improved scalability. While GPUs are capable of dramatically accelerating data intensive tasks, the improved throughput of the SIH pipelines highlights the value of infrastructure-specific optimisation of bioinformatics workflows. Users can process 2,604 samples with the SIH Fastq-to-BAM and 720 samples with the Germline ShortV pipeline in a single day, given a maximum resource allowance of 432 CPU cores and sample batching approach. The Parabricks pipeline allows users to process 195 samples in a single day. The reduced scalability of Parabricks relative to SIH pipelines presented here is a reflection of the number of GPU nodes available at Gadi, resource request limitations imposed by NCI, and limitations of a single license. Parabricks has previously been shown to demonstrate improved scalability with increased GPU resources made available with additional licenses (Franke & Crowgey, 2020). For more details, please see ‘Scalability evaluation methods’ in Supplementary information.

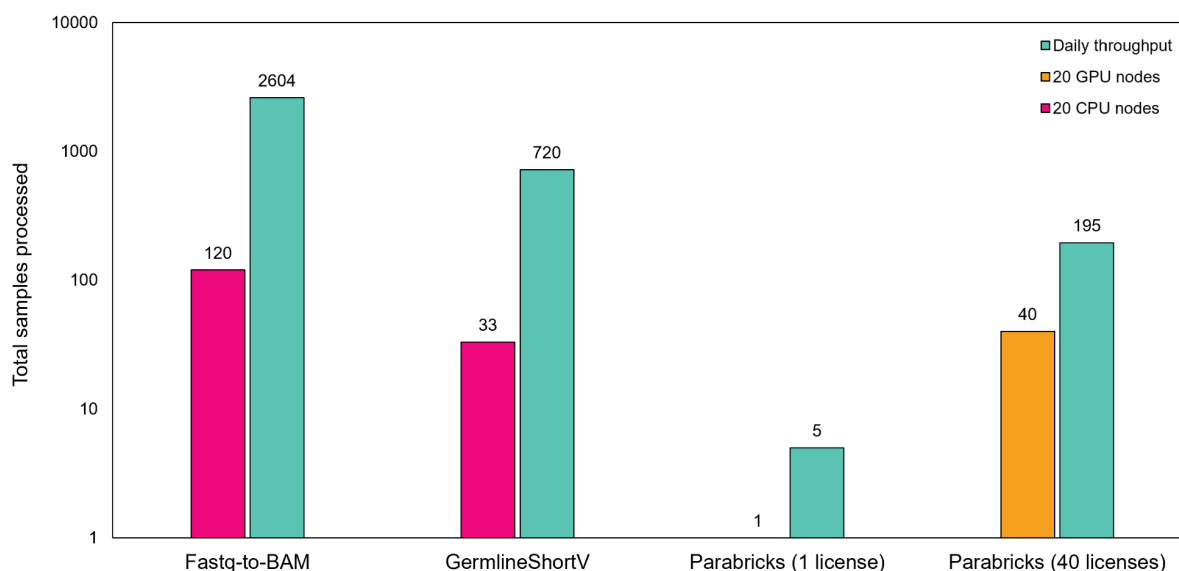


Figure 4: Sample processing capacity of Parabricks and SIH pipelines at NCI Gadi. The number of samples processed with the SIH pipelines, Fastq-to-BAM and GermlineShortV, is reported for 20 CPU nodes (2x24-core Intel Xeon Platinum 8247 [Cascade Lake]) (pink). The number of samples processed with Parabricks is reported for 20 GPU nodes (4x NVIDIA Tesla Volta V100-SXM2-32GB, 2x24-core Intel Xeon Platinum 8247 [Cascade Lake]) (orange). The number of samples processed in one day with the SIH pipeline is reported relative to the maximum resource request limit for the normal queue (432 CPU nodes, 5 hours) (green). The number of samples processed in one day with 1 Parabricks license and within the resource limits of the license (2 GPUs of 1 GPU node), resource request limits of the gpuvolta queue, and is reported (green). The number of samples processed in one day with 40 Parabricks licenses and within the resource limits of the license (20 GPU nodes) is reported (green).

Another essential consideration for these pipelines is their flexibility in meeting the various needs of the Australian bioinformatics community. Best practice recommendations for germline variant calling vary depending on the size of the dataset, clinical relevance, the organism being studied, and the research questions being asked. Both Parabricks and SIH pipelines are modular and can be adjusted by users. For example, this is useful in situations where a researcher is working with an organism that lacks population-level variant datasets and therefore cannot perform base quality score recalibration (BQSR). The SIH pipeline consists of a series of scripts that can be customised according to the user's needs and have been optimised for both human and non-human datasets. Parabricks commands allow users to customise commands through the use of flags that are implemented by the original tools. However, as some tools are packaged up within the same commands by Parabricks, it impedes user's ability to customise the commands for their needs. For example, users wishing to run the read alignment with BWA-MEM in an ALT-aware fashion are unable to do so with Parabricks.

Concluding remarks

We compared the biological accuracy, compute efficiency, and user friendliness of two scalable implementations of WGS best practice pipelines at NCI Gadi: the commercially licensed Parabricks software and the open access SIH Fastq-to-BAM and GermlineShortV pipelines. In addition to biological accuracy, we discuss user friendliness, theoretical ability to scale on NCI's Gadi HPC, flexibility to modify, and the user support model of each pipeline. With default settings, biological benchmarking of resulting SNVs and indels revealed Parabricks and SIH pipelines to have a comparable biological accuracy. Parabricks demonstrated improved recall for both SNVs and indels over the SIH pipeline, however the SIH pipeline demonstrated improved SNV precision. Final callsets were delivered 1.8 times faster by the Parabricks pipeline for the six-sample dataset tested. With access to the maximum available CPU resources on NCI Gadi, the SIH pipelines have the potential to scale processing to 2,604 (Fastq-to-BAM) and 720 (Germline-ShortV) samples per day. The maximum theoretical throughput for Parabricks is 195 samples per day, with access to the maximum available GPU and licenses on Gadi. Scalability of both Parabricks and SIH pipelines are limited by resource request limits, however Parabricks has additional limitations imposed by its licensing model. User preference for one pipeline over the other will depend on the size of their dataset and need to scale, their level of CLI competence, availability of Parabricks licenses, and need to customise the tools used by each pipeline.

Acknowledgements

The work presented in this report was undertaken with resources and services provided by the National Computational Infrastructure (NCI), which is supported by the Australian Government, Parabricks licenses provided by NVIDIA, the Australian BioCommons which is enabled by NCRIS via BioPlatforms Australia funding, and the Sydney Informatics Hub, Core Research

Facility, University of Sydney. We thank Matthew Downton, Javed Shaikh, Ben Menadue, Ankit Sethia, Mehrzad Samadi, Gary Burnett, and Timothy Harkins for their support and feedback.

Author affiliations

1 Sydney Informatics Hub, University of Sydney, Sydney, NSW, Australia

Georgina Samaha, Tracy Chew, Cali Willet & Rosemarie Sadsad

2 Pawsey Supercomputing Centre, Perth, WA, Australia

Sarah Beecroft

3 National Computational Infrastructure, Canberra, ACT, Australia

Brian Davis

References

Australian Genomics Health Alliance (2020). Australian Genomics Data Infrastructure Survey Report: Domestic. October 2020.

AWS Editorial Team, Aniket Deshpande, Olivia Choudhury, Sujaya Srinivasan (2021) Benchmarking the NVIDIA Clara Parabricks germline pipeline on AWS [Blog Post]. Available at: <https://aws.amazon.com/blogs/hpc/benchmarking-the-nvidia-clara-parabricks-germline-pipeline-on-aws/>

Chew, T., Willet, C., Samaha, G., Menadue, B. J., Downton, M., Kobayashi, R., & Sadsad, R. (2021). Germline-ShortV (Version 1.0) [Computer software]. <https://doi.org/10.48546/workflowhub.workflow.143.1>

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernysky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., & Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5), 491–498. <https://doi.org/10.1038/ng.806>

Dos Santos F. R., Buhler, S., Nunes, J. M., Bitarello, B. D., França, G. S., Meyer, D., & Sanchez-Mazas, A. (2015). HLA supertype variation across populations: new insights into the role of natural selection in the evolution of HLA-A and HLA-B polymorphisms. *Immunogenetics*, 67(11-12), 651–663. <https://doi.org/10.1007/s00251-015-0875-9>

Eberle, M. A., Fritzilas, E., Krusche, P., Källberg, M., Moore, B. L., Bekritsky, M. A., Iqbal, Z., Chuang, H. Y., Humphray, S. J., Halpern, A. L., Kruglyak, S., Margulies, E. H., McVean, G., & Bentley, D. R. (2017). A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome research*, 27(1), 157–164. <https://doi.org/10.1101/gr.210500.116>

Franke, K. R., & Crowgey, E. L. (2020). Accelerating next generation sequencing data analysis: an evaluation of optimized best practices for Genome Analysis Toolkit algorithms. *Genomics & informatics*, 18(1), e10. <https://doi.org/10.5808/GI.2020.18.1.e10>

Garcia, M., Juhos, S., Larsson, M., Olason, P. I., Martin, M., Eisfeldt, J., DiLorenzo, S., Sandgren, J., Díaz De Ståhl, T., Ewels, P., Wirta, V., Nistér, M., Käller, M., & Nystedt, B. (2020). Sarek: A portable workflow for whole-genome sequencing analysis of germline and somatic variants. *F1000Research*, 9, 63. <https://doi.org/10.12688/f1000research.16665.2>

Graham, R.L., Shipman, G.M., Barrett, B.W., Castain, R.H., Bosilca, G. and Lumsdaine, A. (2006) "Open MPI: A High-Performance, Heterogeneous MPI," *2006 IEEE International Conference on Cluster Computing*, pp. 1-9, doi: 10.1109/CLUSTER.2006.311904.

Hayes, V., Jaratlerdsiri, W., Jiang, J. et al. (2021) African-specific prostate cancer molecular taxonomy, 01 December 2021, PREPRINT (Version 1)
<https://doi.org/10.21203/rs.3.rs-1122619/v1>

Jarlier, F., Joly, N., Fedy, N. et al. (2020) QUARTIC: QUick pARallel algoRithms for high-Throughput sequencing data proCessing [version 3; peer review: 2 approved]. *F1000Research*, 9(240) (<https://doi.org/10.12688/f1000research.22954.3>)

Krusche, P., Trigg, L., Boutros, P. C., Mason, C. E., De La Vega, F. M., Moore, B. L., Gonzalez-Porta, M., Eberle, M. A., Tezak, Z., Lababidi, S., Truty, R., Asimenos, G., Funke, B., Fleharty, M., Chapman, B. A., Salit, M., Zook, J. M., & Global Alliance for Genomics and Health Benchmarking Team (2019). Author Correction: Best practices for benchmarking germline small-variant calls in human genomes. *Nature biotechnology*, 37(5), 567.
<https://doi.org/10.1038/s41587-019-0108-0>

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16), 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>

Li H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics (Oxford, England)*, 27(21), 2987–2993.
<https://doi.org/10.1093/bioinformatics/btr509>

Lin M.F., Roden, O., Penn, J., Bai, X., Reid, J.G., Krasheninina, O., Salerno, W.J. (2018) GLnexus: joint variant calling for large cohort sequencing. *bioRxiv*, 343970. doi: 10.1101/343970.

- Lott, M. J., Wright, B. R., Neaves, L. E., Frankham, G. J., Dennison, S., Eldridge, M. D. B., Potter, S., Alquezar-Planas, D. E., Hogg, C. J., Belov, K., & Johnson, R. N. (2022). Future-proofing the koala: Synergising genomic and environmental data for effective species management. *Molecular Ecology*, 00, 1–21. <https://doi.org/10.1111/mec.16446>
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9), 1297–1303. <https://doi.org/10.1101/gr.107524.110>
- NVIDIA (2020). NVIDIA Clara-Parabricks Pipelines (Version 3.5) https://docs.nvidia.com/clara/parabricks/v3.5/text/software_overview.html#nvidia-clara-parabricks-pipelines
- Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M. R., Zschocke, J., & Trajanoski, Z. (2014). A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in bioinformatics*, 15(2), 256–278. <https://doi.org/10.1093/bib/bbs086>
- Pérez-Wohlfeil, E., Torreno, O., Bellis, L. J., Fernandes, P. L., Leskosek, B., & Trelles, O. (2018). Training bioinformaticians in High Performance Computing. *Heliyon*, 4(12), e01057. <https://doi.org/10.1016/j.heliyon.2018.e01057>
- Satgunaseelan, L., Porazinski, S., Strbenac, D., Istadi, A., Willet, C., Chew, T., Sadsad, R., Palme, C. E., Lee, J. H., Boyer, M., Yang, J., Clark, J. R., Pajic, M., & Gupta, R. (2021). Oral Squamous Cell Carcinoma in Young Patients Show Higher Rates of *EGFR* Amplification: Implications for Novel Personalized Therapy. *Frontiers in oncology*, 11, 750852. <https://doi.org/10.3389/fonc.2021.750852>
- Willet, C., Chew, T., Samaha, G., Menadue, B. J., Downton, M., Kobayashi, R., & Sadsad, R. (2021). Fastq-to-BAM (Version 2.0) [Computer software]. <https://doi.org/10.48546/workflowhub.workflow.146.1>
- Yun, T., Li, H., Chang, P. C., Lin, M. F., Carroll, A., & McLean, C. Y. (2021). Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *Bioinformatics (Oxford, England)*, 36(24), 5582–5589. Advance online publication.
- Zhao, S., Agafonov, O., Azab, A. *et al.* (2020). Accuracy and efficiency of germline variant calling pipelines for human genome data. *Scientific Reports* 10, 20222. <https://doi.org/10.1038/s41598-020-77218-4>

Supplementary information

Platinum Genomes dataset

Whole genome sequencing data for six of the 17 member, 'Platinum Genomes' pedigree was used for this report. These samples comprised two mother-father-offspring trios (Figure S1) and were developed as benchmarking datasets by Illumina. All samples were sequenced to 50x depth on an Illumina HiSeq 2000 system (Table S1).

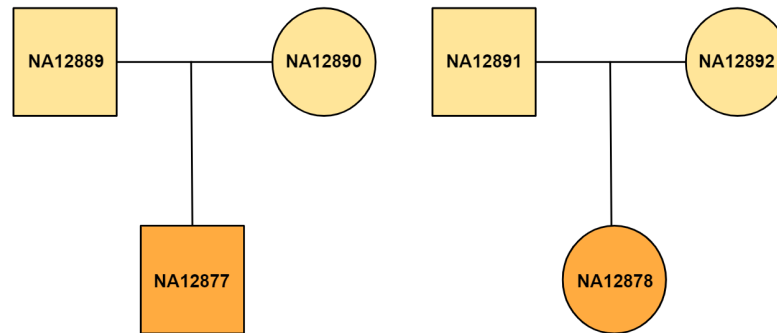


Figure S1: Platinum Genomes mother-father-offspring trios used in this report.

Sequencing data for the remaining pedigree members were not consented for public release. Fastq files for all samples were downloaded from Illumina's BaseSpace Sequence Hub, along with short variant truth sets in VCF format for each sample and high confidence region files. Variant truth sets were generated from a genome-wide catalog of 5.4 million phased variants reported from six different variant calling pipelines (Eberle et al. 2016). The Platinum Genomes samples provide a highly accurate, comprehensive, genome-wide set of variants that have previously been used to improve variant calling algorithms, and develop Best Practice recommendations for germline variant calling methods.

Table S1: Fastq file details for each Platinum Genomes sample downloaded from Illumina's BaseSpace Sequencing Hub. Mapping coverage is given for each sample for Parabricks and SIH pipelines.

SampleID	Fq.gz size (GB)	Total reads	Final bam size (GB)	Parabricks average coverage	SIH average coverage
NA12877	155	1,637,816,924	146	50.0903	51.5995
NA12878	146	1,586,092,978	138	48.3747	49.8143
NA12889	163	1,734,633,792	160	52.7796	54.3713
NA12890	131	1,428,359,024	124	43.3076	44.4889

NA12891	144	1,571,604,064	136	47.2721	48.7023
NA12892	156	1,702,988,434	147	51.6729	53.2171

Pipeline details

We compared the performance of two genome alignment and germline variant calling pipelines run on the National Computational Infrastructure's (NCI) Gadi supercomputer. The pipelines were a commercial GPU implementation developed by NVIDIA Parabricks and an open source CPU implementation developed collaboratively by bioinformaticians at the Sydney Informatics Hub and HPC specialists at NCI. Both of these pipelines have implemented GATK best practice methods for short read mapping and germline variant calling. A summary of the software used to perform each step by both pipelines is provided in Table S2.

Table S2: Steps and software used in the Parabricks and SIH pipelines evaluated within this document.

Step	Parabricks	SIH
Read alignment	GPU-BWA mem, Parabricks sort reads*	BWA kit with ALT contig post-processing
Mark duplicates	Parabricks Mark Duplicates*	Samblaster to mark duplicates, remove discordant and split reads
Base quality score recalibration (BQSR)	Parabricks BQSR*	GATK BaseRecalibrator, SplitIntervals, GatherReports
Apply BQSR	Parabricks applybqsr	GATK ApplyBQSR, GatherBamFiles
Variant calling	Parabricks HaplotypeCaller	GATK HaplotypeCaller, GatherVCFs
Joint variant calling	Parabricks GLnexus	GenomicsDBImport, GATK GenotypeGVCFs, GatherVCFs
Variant quality score recalibration (VQSR)	NA	GATK VariantFiltration, VariantRecalibrator, ApplyVQSR, CollectVariantCallingMetrics

* Processes run by Parabricks 'fq2bam' command.

Pipeline evaluation methods

Mapping and germline variant calling pipelines developed by NVIDIA Parabricks and SIH were run on six Platinum Genomes individuals according to their developer instructions on NCI Gadi. Pipelines were evaluated on their biological accuracy, compute performance, and user friendliness.

Four commands were run from the Parabricks v3.5 suite to generate BAM, individual gVCF and cohort BCF outputs given a pair of fastq files. Each command was run with 2 GPUs, however fq2bam, haplotypcaller and gl nexus tools are all capable of scaling beyond 2 GPUs. The Parabricks pipeline 'fq2bam' command aligned paired-end reads to the Hg38 reference genome, and sorted reads based on their genomic position, after marking duplicates and generating base quality score recalibration (BQSR) report. The 'applybqsr' command was run to update base quality scores using the BQSR report for each sample. Each sample's final bam file was then input to the 'haplotypcaller' command to generate gVCF files for all sites in the Hg38 assembly. Joint genotyping of all samples was performed by running the 'gl nexus' command, generating the final cohort variant set in BCF format.

Two SIH pipelines were run to perform mapping (Fastq-to-BAM) and germline variant calling (Germline-ShortV) of six Platinum Genomes samples. These pipelines were developed by bioinformaticians at the Sydney Informatics Hub in collaboration with HPC specialists at NCI, specifically for the Gadi supercomputer and are implementations of the BROAD Institute's Best Practice workflow. Both the alignment and germline variant calling pipelines consist of a series of bash scripts to be run manually by the user and leverage multiple nodes to run all stages of the workflow in parallel. Each step was run according to the resource recommendations provided in the pipeline documentation which is available on their GitHub repository (<https://github.com/Sydney-Informatics-Hub/Bioinformatics>) along with the scripts.

Genome coverage of BAM files produced by Parabricks and SIH pipelines was determined using Samtools depth (Li 2009). Total variants, SNVs, indels, and transition and transversion ratios were collected using Bcftools stats (Li 2011) for each pipeline's cohort-level VCF file. Biological accuracy metrics of precision, recall, and F1-score were determined using the haplotype VCF comparison tool, Hap.py (Krusche et al. 2019), relative to a truth set of variants. Recall, precision, F1 scores were calculated by Hap.py as:

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

$$\text{F1-score} = 2 * \text{Precision} * \text{Recall} / (\text{Precision} + \text{Recall})$$

Where TP is the number of 'true positives' or variants called that match the truth set, FP is the number of 'false positives' or variants with mismatching genotypes or alt-allele in regions the truthset called as homozygous-reference, and FN is 'false negatives' or the variants present in the truth set that failed to be called. Sample-level truth sets were downloaded with raw data in vcf format from Illumina BaseSpace Sequencing Hub, for all six Platinum Genomes samples

(<https://basespace.illumina.com/s/2K7LqNG7Mt1h>). Technical performance metrics of walltime, CPU/GPU and RAM usage, and efficiency were collected using HPC usage report scripts, developed by the Sydney Informatics Hub (https://github.com/Sydney-Informatics-Hub/HPC_usage_reports).

Scalability evaluation methods

Potential scalability of Parabricks and SIH pipelines were captured considering the resource request limitations imposed by NCI Gadi (<https://opus.nci.org.au/display/Help/Queue+Limits>). SIH pipelines were run on the normal and normalbw queues designed for standard computationally intensive jobs. Users can request a maximum of 432 2x24-core Intel Xeon Platinum 8247 (Cascade Lake) nodes on the normal queue. Parabricks was run using the gpuvolta queue which consists of 160 4x NVIDIA Tesla Volta V100-SXM2-32GB and 2x24-core Intel Xeon Platinum 8247 (Cascade Lake) nodes. The most computationally intensive tasks of 'align', 'apply_bqsr', and 'haplotypcaller' were the limiting processes of the SIH pipelines and all run on the normal queue. As such we reported the scalability limitations of the SIH pipeline considering the limits of the normal queue at NCI Gadi.

Considering differences in availability of CPU and GPU nodes at NCI Gadi, three comparisons were made to evaluate the scalability of Parabricks and SIH pipelines:

1. The maximum number of samples processed by each pipeline in parallel. This comparison considers resource request limitations imposed by NCI Gadi. For the normal queue (and SIH pipelines) this is 432 CPU nodes for 5 hours and for the gpuvolta queue (and Parabricks pipeline) this is 20 GPU nodes for 5 hours.
2. The differences in scalability offered by 20 GPU nodes for the Parabricks pipeline and 20 CPU nodes for the SIH pipelines.
3. The maximum daily sample throughput of each pipeline, permitted by batch processing of the maximum number of samples processed in parallel at one time.

The resource requirements and sample processing capacity for scalability comparisons are presented below.

Table S3.1 Maximum number of samples processed by each pipeline, considering normal and gpuvolta queue request limitations. The maximum number of samples processed by Parabricks does not take licensing restrictions (2GPUs/license) into account.

Pipeline	Queue	Max node request	Max walltime (mins)	Total walltime (mins)	Number of batches /day	Max samples /1 run	Max samples /day
Parabricks	gpuvolta	20	300	294	4.89	40	195
SIH Fastq-to-BAM	normal	432	300	279.1	5.15	506	2604
SIH GermlineShortV	normal	432	300	288	5	144	720

Table S3.2 Maximum number of samples processed by 20 GPU nodes and 20 CPU nodes.

Pipeline	Queue	Max node request	Max walltime (mins)	Total walltime (mins)	Max samples
Parabricks	gpuvolta	20 GPU	300	294	40
SIH Fastq-to-BAM	normal	20 CPU	300	279.1	120
SIH GermlineShortV	normal	20 CPU	300	288	33