Comparing Form Bio Data Analysis Performance With and Without NVIDIA Clara Parabricks

Brandi Cantarel, PhD
Director of Bioinformatics, Form Bio

Abstract

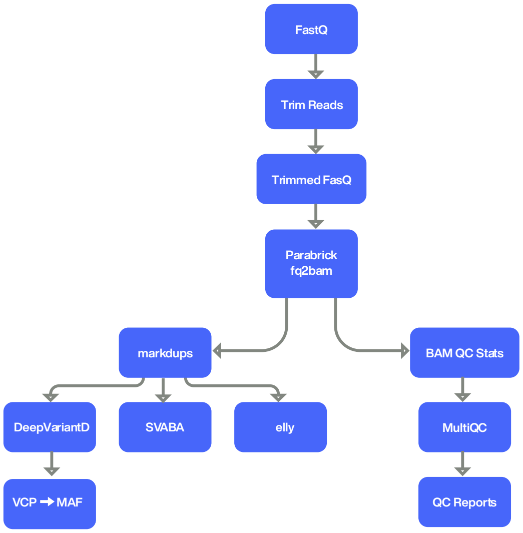
Genetic variant detection has wide application in genomics including identifying genetic changes that drive phenotypic changes. The detection of differences between a genomic re-sequencing experiment allows researchers to detect genetic diversity with a population, identify disease causing genetic changes and discover genetic markers of speciation when comparing closely related species. Germline variant detection is used to detect inherited genetic changes, whereas somatic variant detection is used to detect acquired genetic changes including the genetic mutations that drive tumor growth. There are many types of genetic variation including single nucleotide variants (SNVs) and insertions/deletions (indels), copy number variants (CNVs), and structural variants (SVs), which include large insertions, deletions, inversions, duplications and translocations. Over the last decade several open source tools have been developed to improve the accuracy and speed of SNV/Indel variant detection including BWA mem, minimap2, GATK and DeepVariant. However it can still take days or weeks with these tools to analyze whole genome sequencing without very large computational clusters so that analysis runs in parallel over many machines. Accelerated versions of these tools using GPUs has been shown to reduce computational time and resources. We analyzed a publicly available dataset using DNA from a sample named Genome in a Bottle (GIAB), NA12878 to determine the sensitivity and specificity of running NVIDIA Parabricks compared to the open source versions of the same tools on Google Cloud Platform. The resulting analysis using NVIDIA Parabricks was faster and cheaper than using open source tools. The optimization did not significantly change the underlying alignments produced. Sensitivity and specificity were similar to the open source version of the tools even when only the alignment was used.

Methods

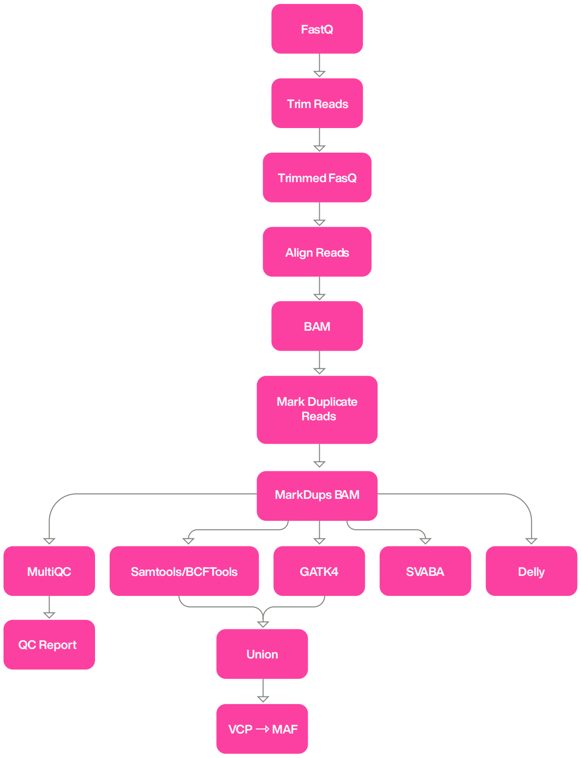
The workflow “Genomics: Germline Variant Analysis” determines genetic variants including SNVs, insertions and deletions of high-quality NGS data compared to a reference genome. Reads are trimmed using TrimGalore[1–3], to trim low quality (qual < 25) ends of reads and remove reads < 35bp. This workflow can be run with native open-source tools (NOST) or with Parabricks.

With NOST, trimmed reads are aligned to a reference genome using BWA mem[4] or Minimap[25]. Duplicates reads can optionally be marked using Picard MarkDuplicates[6]. BAMs from the same sample generated by multiple runs are merged using samtools[7]. Alignment qualtity is assessed using FastQC2, Samtools[7], Bedtools[8]. With Parabricks, trimmed reads are aligned, duplicate reads are marked and alignment quality is accessed using fq2bam[9,10]. Quality metrics are summarized with MultiQC. With NOST, variants can be detected with joint calling using Freebayes[11], Samtools/Bcftools[7] and GATK4[12]. With Parabricks, variants can be detected with GATK[12] and DeepVariant[13] to produce gVCF files. Genotyping of gVCF files is determined using[14]. Variants effects are determined using SNPEff[15].

Workflow Diagrams



With NVIDIA Parabricks



Native open-source tools

Results

We ran our workflow using open source versions of alignment and variant calling and with Parabricks with 3 separate sequencing runs of NA12878 as separate samples spanning 20M-160M reads. The full workflow ran in 344 minutes using Parabricks and 2,793 minutes using open source tools with multiple steps running it parallel, representing an 88% time savings. The biggest increase in efficiency came from variant calling where the median job took 1,376 minutes for open source tools and 94 minutes using Parabricks, representing a 93% time savings. The cost of running with Parabricks was US$45 vs US$93 with open source tools, representing a 52% savings. The resulting variants have comparable sensitivity, within 3%, for variant calling methods that use Parabricks including GATK and DeepVariant and those that do not have an accelerated version such as Freebayes, Samtools.

SAMPLE ID	METHOD	SN	PPV
NOST	gatk	95.2	94.8
NP	gatk	92.2	93.8
NOST	deepvariant	91.0	98.9
NP	deepvariant	87.8	99.6

References

1. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. journal 17, 10 (2011).
2. Andrews, S. et al. FastQC. (2012).
3. Krueger, F., James, F., Ewels, P., Afyounian, E. & Schuster-Boeckler, B. FelixKrueger/TrimGalore: v0.6.7 - DOI via Zenodo. (2021) doi:10.5281/ZENODO.5127899.
4. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv Prepr. ArXiv 00, 3 (2013).
5. Li, H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094–3100 (2018).
6. Thomer, A. K., Twidale, M. B., Guo, J. & Yoder, M. J. Picard Tools. in Conference on Human Factors in Computing Systems - Proceedings (2016).
7. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).
8. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010).
9. Franke, K. R. & Crowgey, E. L. Accelerating next generation sequencing data analysis: an evaluation of optimized best practices for Genome Analysis Toolkit algorithms. Genomics Inform. 18, e10 (2020).
10. Friedman, S., Gauthier, L., Farjoun, Y. & Banks, E. Lean and deep models for more accurate filtering of SNP and INDEL variant calls. Bioinformatics 36, 2060–2067 (2020).
11. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. (2012).
12. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43, 491–498 (2011).
13. Poplin, R. et al. A universal SNP and small-indel variant caller using deep neural networks. Nat. Biotechnol. 36, 983–987 (2018).
14. Lin, M. F. et al. GLnexus: joint variant calling for large cohort sequencing. http://biorxiv.org/lookup/doi/10.1101/343970 (2018) doi:10.1101/343970.
15. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin) 6, 80–92 (2012).