SPRITE: A Fast Parallel SNP Detection Pipeline

Kamesh Madduri, Vasudevan Rengasamy, Paul Medvedev

The Pennsylvania State University

sprite-psu.sourceforge.net

Abstract

Variant calling analysis of whole genome sequencing data is a computationally-intensive process, often taking several hours to even multiple days. We develop SPRITE, an open-source toolkit for detecting Single Nucleotide Polymorphisms (SNPs). By combining fast new algorithms and parallel computing technologies, SPRITE significantly accelerates the computational tasks of alignment, intermediate file processing, and SNP detection. For a benchmark human genome with 30X coverage, SPRITE takes just 30 minutes (from FASTQ data ingestion to VCF file creation) on a compute cluster with 16 servers. We also evaluate SPRITE on several Illumina Platinum Genome data sets. PARSNIP, the SNP detection tool in SPRITE, is orders-of-magnitude faster than the SNP callers in GATK, samtools, and Freebayes. In our preliminary analysis, we find that that the quality of results obtained (PARSNIP precision and recall using high-confidence variant calls as ground truth) is comparable to state-of-the-art SNP-calling software. A prototype implementation of SPRITE is available at sprite-psu.sourceforge.net.

SPRITE Overview

Single Nucleotide Polymorphisms (SNPs) are the most studied type of structural variation. SNPs are nucleotide differences at a single position in the genome. SPRITE [1] aims at end-to-end acceleration of the SNP detection workflow.

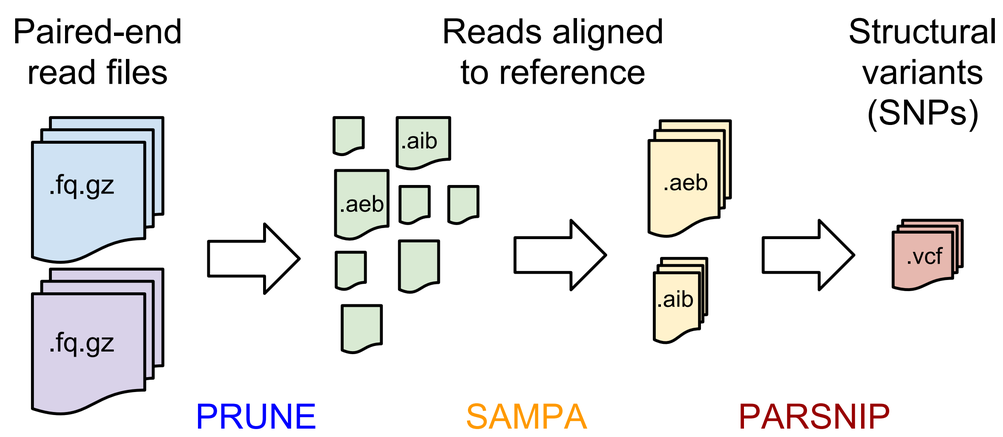

Figure: A simplified view of computational stages in a SNP detection pipeline. We indicate the I/O formats for the three new tools comprising the SPRITE pipeline.

- PRUNE is currently based on the BWA-MEM algorithm. For parallelism, the read files are partitioned across compute nodes, and the reference sequence is replicated on every node.
- We introduce a new intermediate binary format (.aeb, .aib) for alignment output. SAM file creation is optional.
- SAMPA performs a parallel sort of alignment output and prepares data in a binary format that is amenable to SNP calling.

PARSNIP

PARSNIP is a simple counting-based SNP detection tool. It reads SAMPA output to update a nucleotide frequency table $\mathcal{F}$. Compute nodes concurrently process reference contigs for parallel execution.

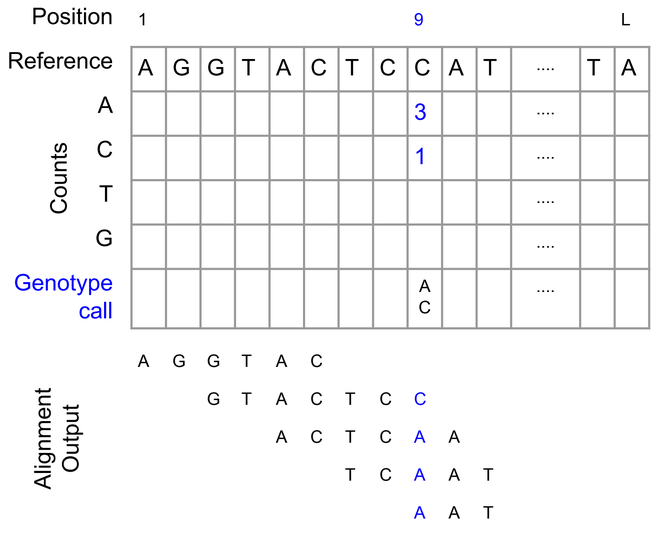

Figure: Organization of the frequency table $\mathcal{F}$ used in PARSNIP.

SPRITE and PARSNIP Evaluation

- SPRITE is a collection of command-line tools. We use the C programming language with MPI and Pthreads libraries for parallelism.
- We evaluate SPRITE performance on Illumina platinum genome sequence data (NA12878 sequenced to 50x depth on an Illumina HiSeq 2000 system) and data from the SMaSH variant detection benchmarking toolkit.
- We compare performance to a 'reference' pipeline using BWA (v0.7.12), SAMtools (v1.1), GATK (v3.2.2).
- For SNP-calling, we compare PARSNIP output to the high-confidence variant calls for NA12878 provided by the Platinum Genomes project.
- On 16 nodes of the NERSC Edison system (2 12-core Intel Ivy Bridge processors and 64 GB memory per node), SPRITE takes 30 minutes to process the synthetic Venter data set from SMaSH.
- Performance results on NA12878 and TACC Stampede supercomputer are presented below. Each compute node has two 8-core Intel Sandy Bridge processors and 32 GB memory. We do not use the Intel Xeon Phi accelerators.

Pipeline Stage	Ref. Pipeline, 1 node Tool	Time (min)	SPRITE, 1 node Time (min)	Speedup	SPRITE, 16 nodes Time (min)	Speedup
Alignment	BWA-MEM	580	692	0.8×	54.6	10.6×
SAM file processing	SAMtools	526	60	8.7×	7.0	75.1×
SNP Calling	GATK-UG	63	12	5.2×	3.0	21.0×
Overall		1169	764	1.5×	64.6	18.1×

Table: End-to-end pipeline execution times and speedup on Stampede for Illumina NA12878 data set. BWA-MEM, SAMtools, and GATK only support single-node (16 cores) parallelism.

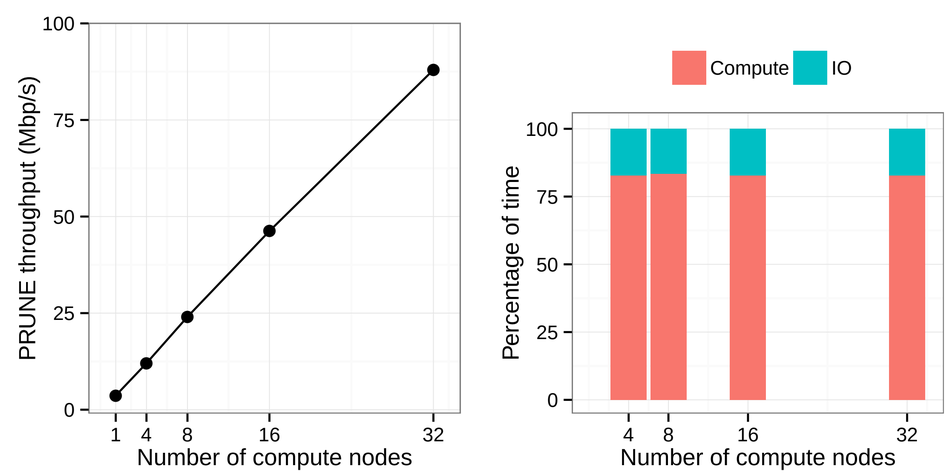

Figure: PRUNE scales almost linearly with the number of compute nodes on Stampede. Both Compute and IO time reduce with increasing parallelism.

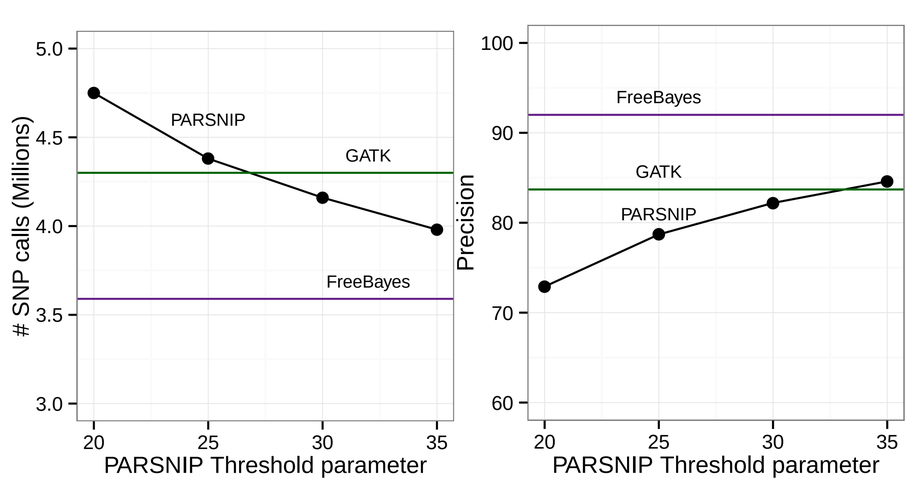

Figure: PARSNIP's SNP count and precision depend on a user-configurable threshold parameter.

Tool	Quality Precision	Recall	Running time (min)
PARSNIP	78.7	93.9	12
FreeBayes (v0.9.21)	92.0	90.0	146
GATK (v3.2.2)	83.7	98.1	63

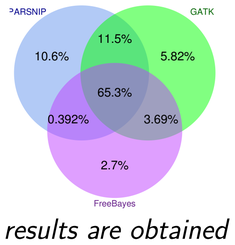

Figure: SNP-calling quality comparison on NA12878. PARSNIP results are obtained by setting the threshold parameter to 25.

References

[1] V. Rengasamy, K. Madduri, Engineering a high-performance SNP detection pipeline, Penn State Computer Science and Engineering Technical Report, April 2015.

SPRITE: A Fast Parallel SNP Detection Pipeline