Genomic Workflow Acceleration on Supercomputers

Kamesh Madduri, Mahmut Kandemir,
Paul Medvedev, Padma Raghavan

The Pennsylvania State University

NSF XPS Workshop
June 2, 2015

sites.psu.edu/XPSGenomics

Next-generation sequencing (NGS) data analyses workflows

+

~ 250 GB
“raw data”

Quality Assessment
Filtering, trimming

Read Alignment
Given reference genome or ref. sequences

De novo assembly

Variant Identification
SNP, Indel, CNV, SV

~ 10 MB

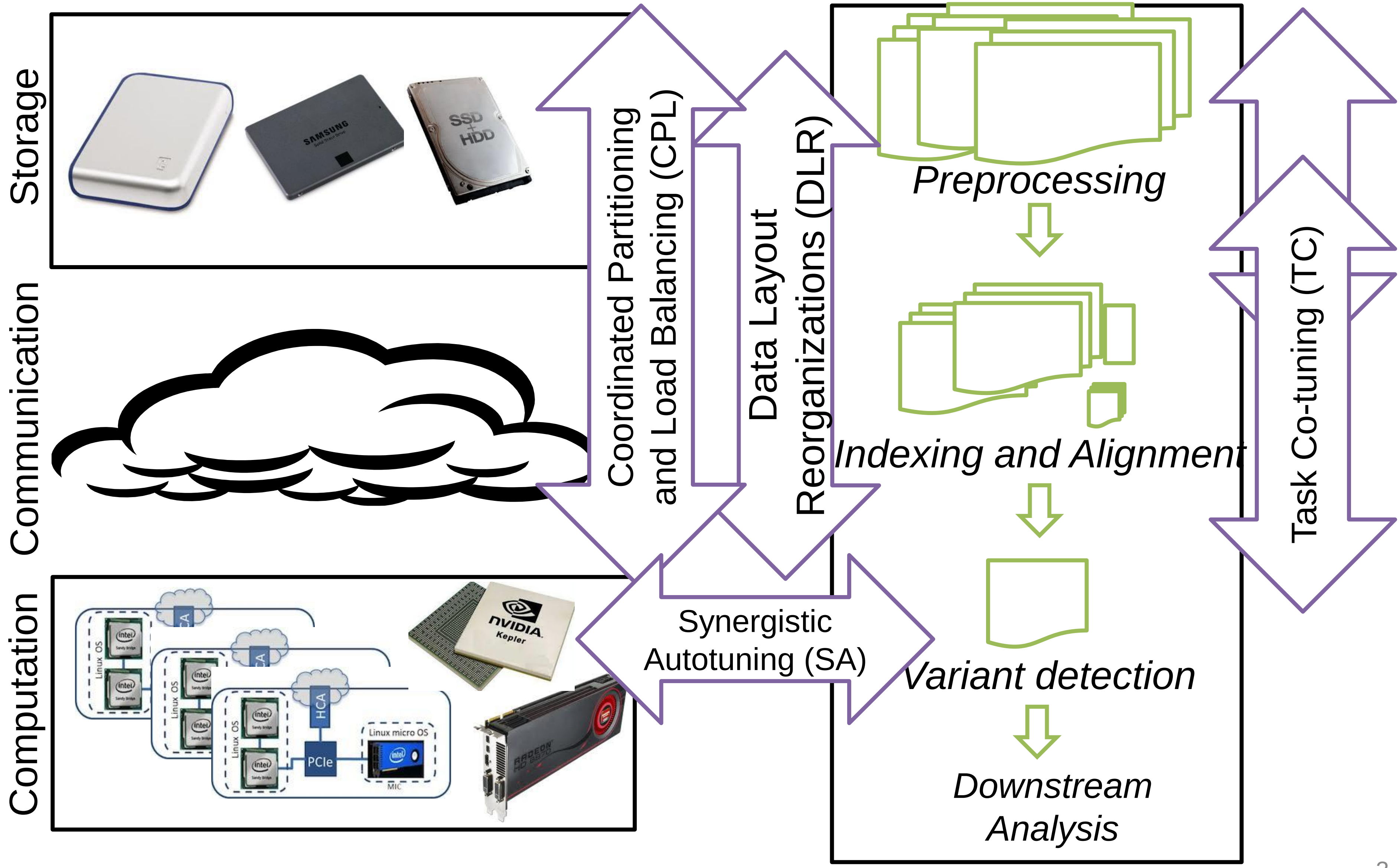
Annotation
public databases

Visualization
Variants, annotation

Prioritization, Filtering
Downstream analysis

Lab validation

Our XPS Project: Accelerating the **genetic variant detection** workflow on **(heterogeneous) supercomputers**



Our XPS Project: Accelerating the **genetic variant** detection workflow on **(heterogeneous) supercomputers**

~ 250 GB sequencing data

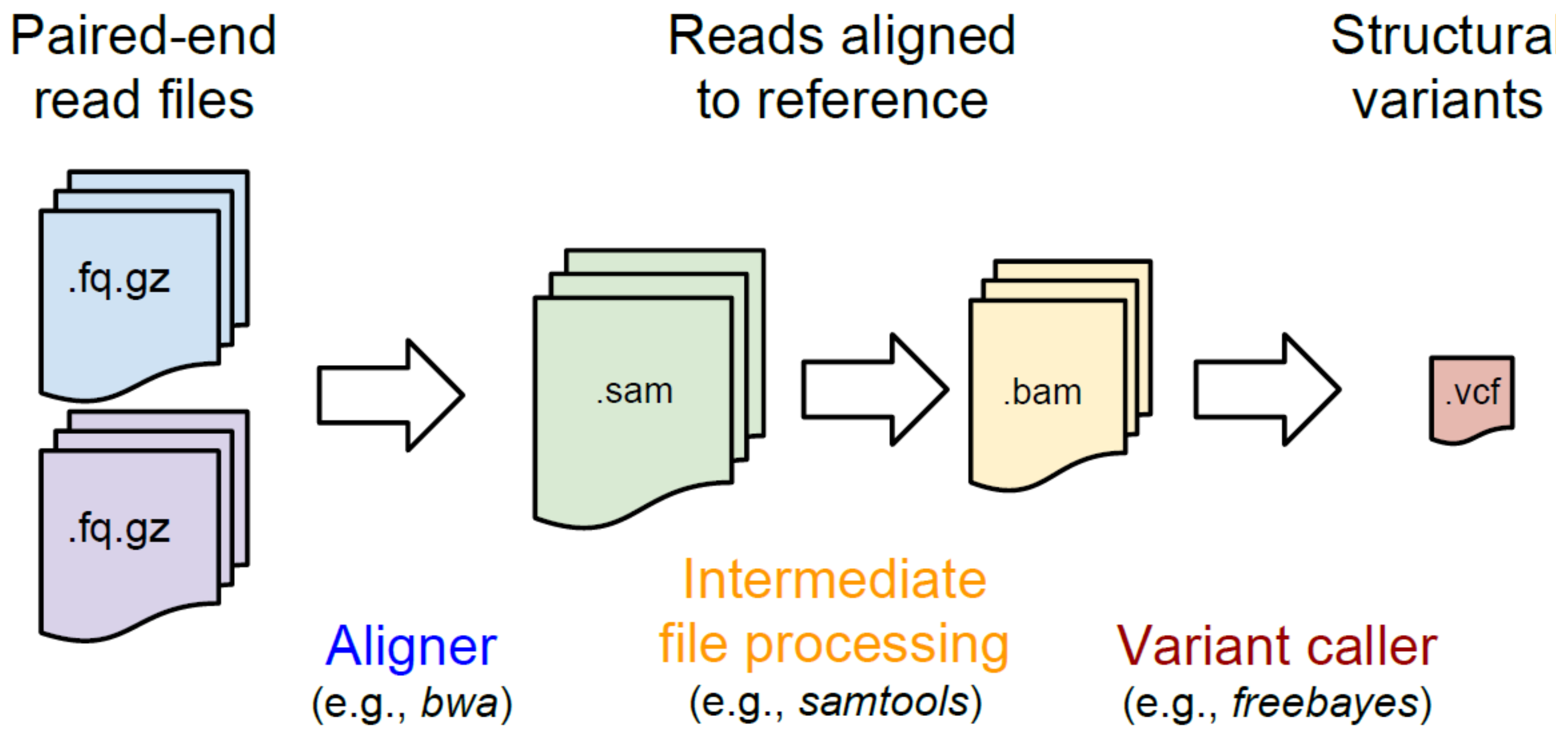
Takes **28 hours** on a workstation using state-of-the-art tools.

We target end-to-end optimizations with modest parallelism + new parallel algorithms, accelerators, and parallel I/O tuning.

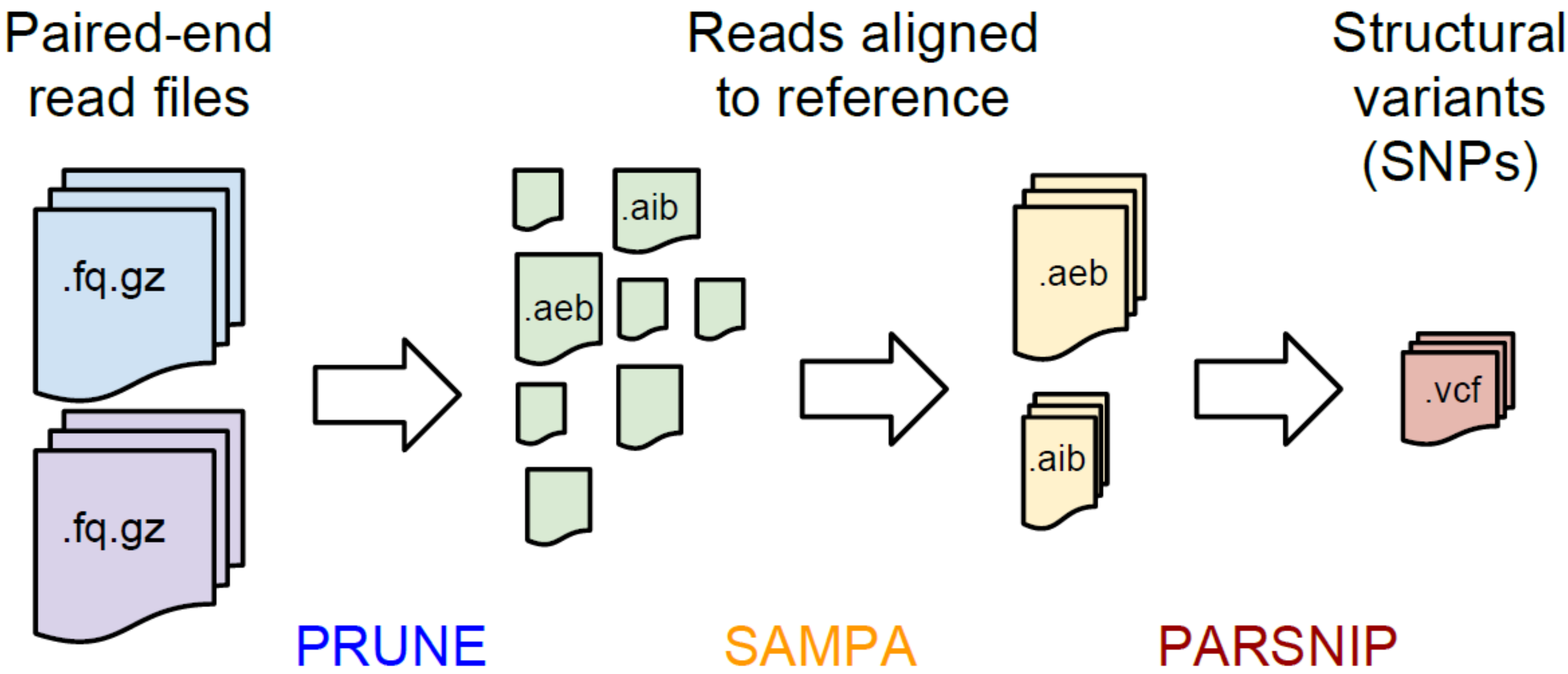
~ 10 MB structural variants

SPRITE: HPC pipeline for SNP detection

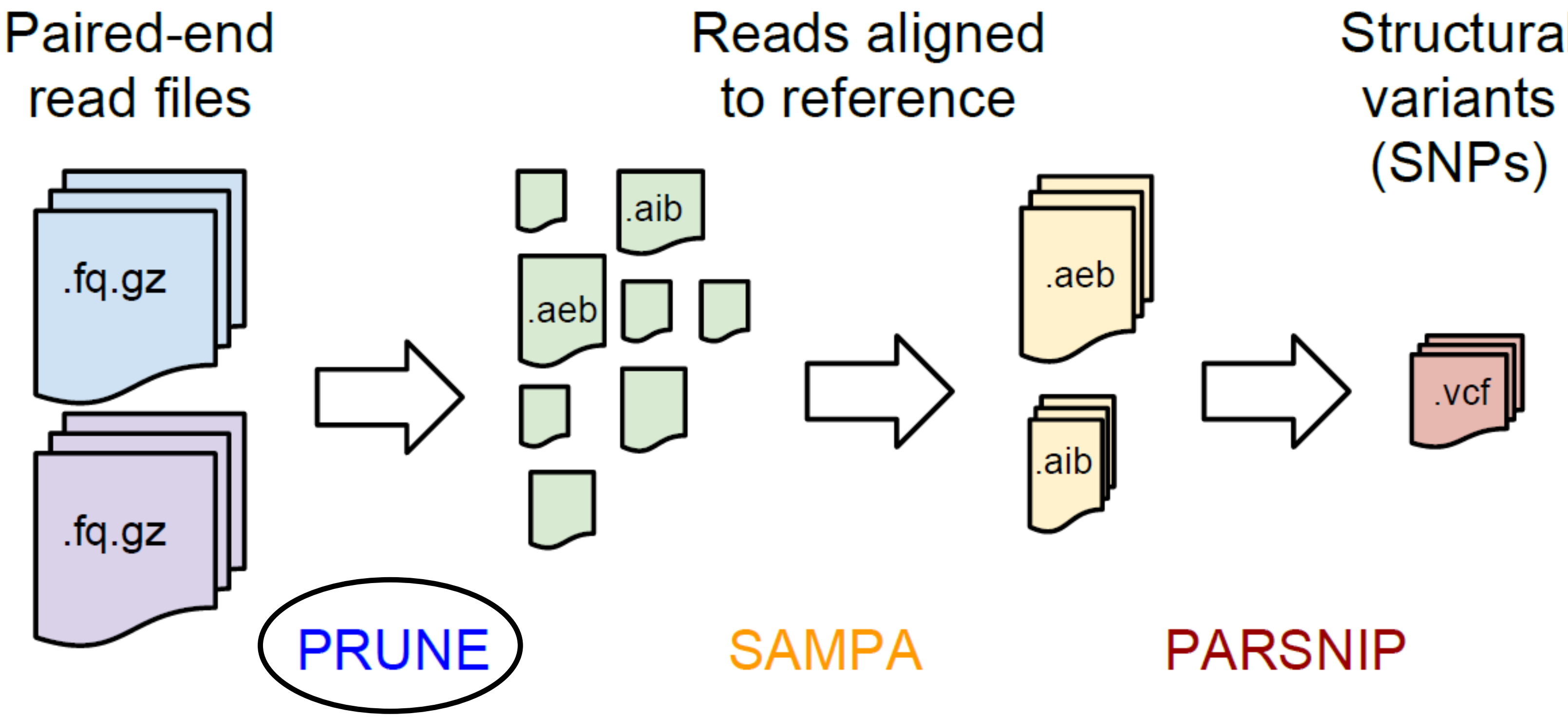
A "reference" pipeline



SPRITE: HPC pipeline for SNP detection

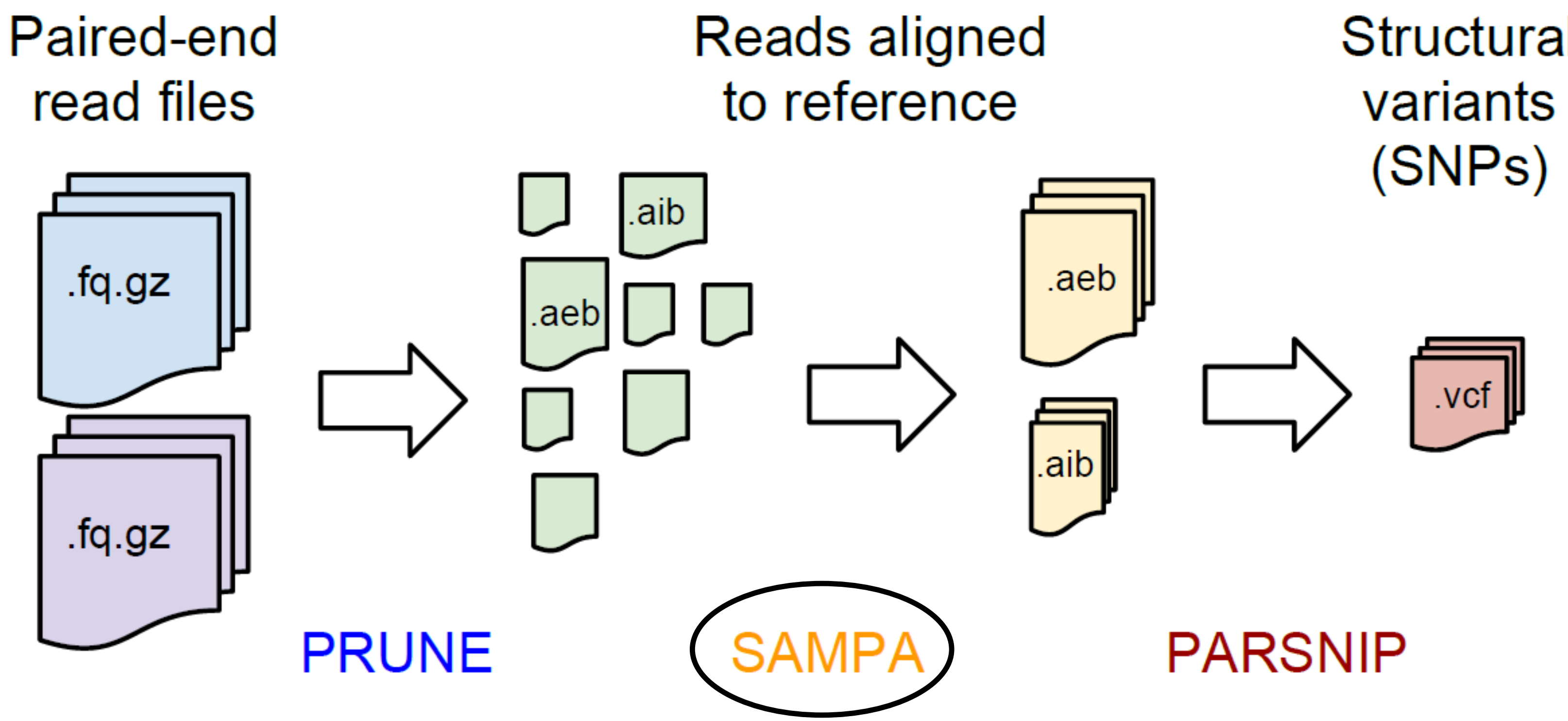


SPRITE: HPC pipeline for SNP detection



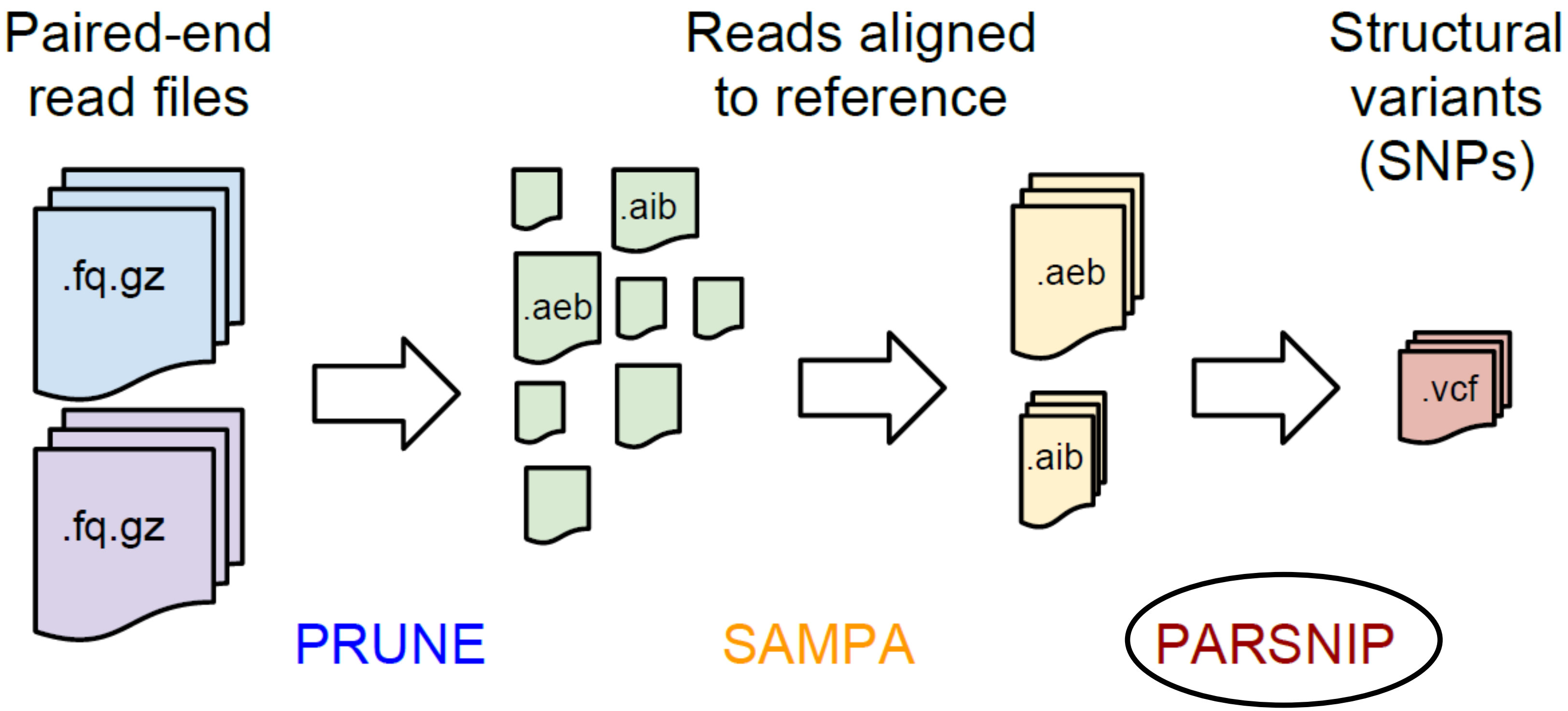
Based on BWA-MEM
Changes to support multi-node parallelism

SPRITE: HPC pipeline for SNP detection



New tool
Parallel in-memory sort
Operates on binary intermediate files

SPRITE: HPC pipeline for SNP detection



New tool
Optimizations for the common case
Multigrained parallelism

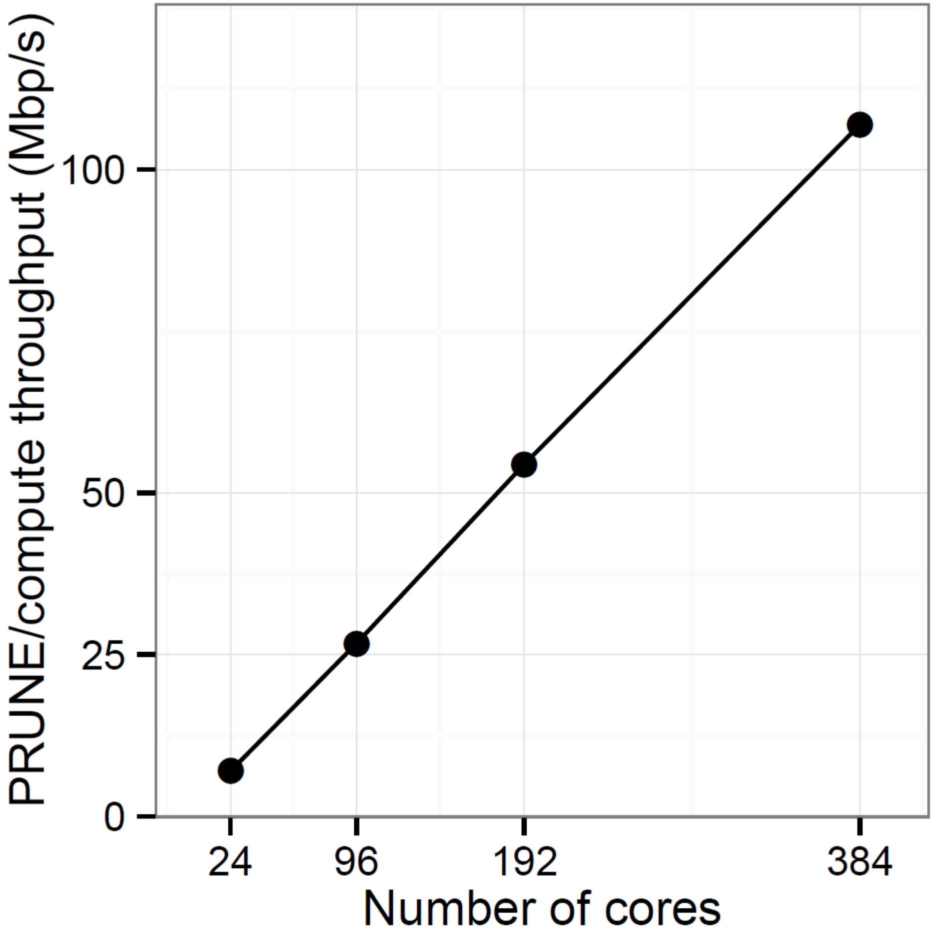
Result: Intermediate steps are no longer the bottleneck

- SNP detection workflow on SMaSH Venter data set.
- 16 nodes of NERSC Edison supercomputer. Each node has two Intel 12-core Ivy Bridge processors and 64 GB memory. Lustre shared file system with 72 GB/s peak I/O performance.
- Reference pipeline takes 28 hours on one node
- SPRITE: 31 minutes on 16 nodes

Result: Intermediate steps are no longer the bottleneck

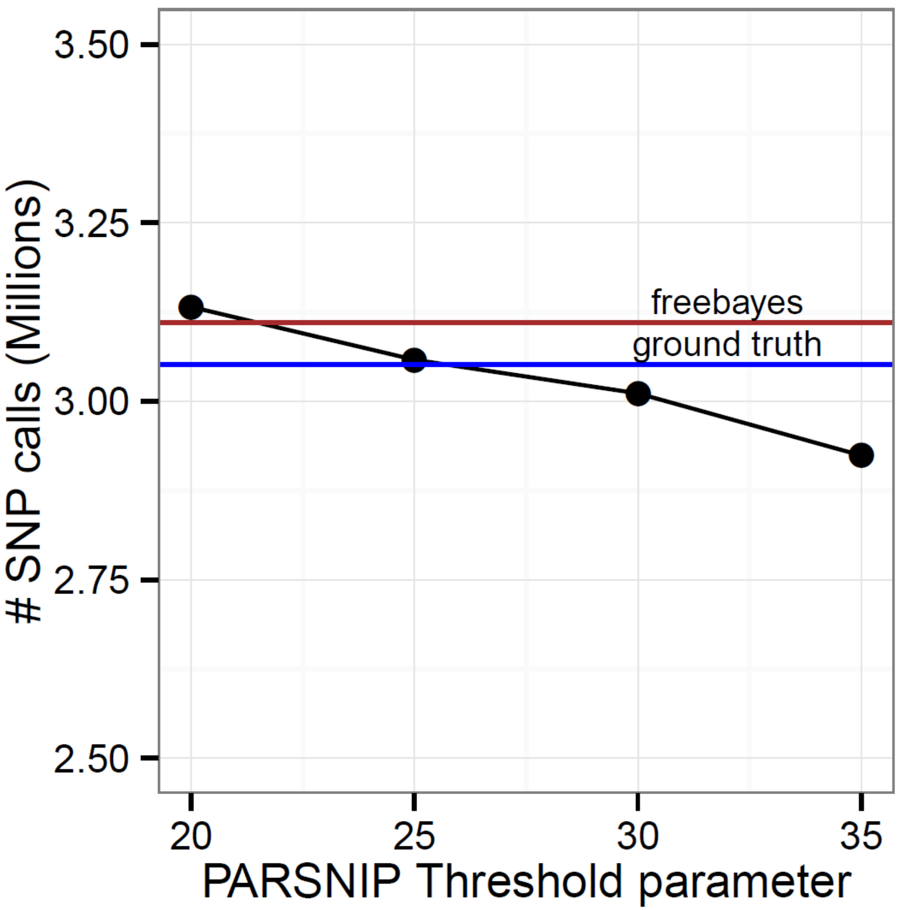
Pipeline	Ref. Pipeline, 24 cores		SPRITE, 384 cores	
Stage	Tool	Time (min)	Time (min)	Speedup
Alignment	bwa	393	26.36	14.91×
SAM file processing	samtools	401	3.40	117.94×
SNP Calling	freebayes	889	1.55	573.55×
Overall		1683	31.31	53.75×

Result: Compute phases in Alignment scale reasonably well



Result: SNP detection quality using **PARSNIP** is comparable to state-of-the-art tools

Tool	Precision	Recall
PARSNIP	95.1	97.2
freebayes	94.8	97.2
mpileup	98.7	97.0
GATK	99.3	91.7



Ongoing work targeting the variant detection workflow

- Tuning SPRITE for alternate hardware configurations.
- Lightweight in-memory data layout reorganizations.
- Avoiding I/O in intermediate steps.
- Alternate intermediate and output representations.

Cross-cutting

- PARSNIP GPU and Xeon Phi parallelization.
- Parallel tools for structural variant detection.
- Adding probabilistic models to PARSNIP.

Variant detection

- I/O Optimizations in alignment step.
- Alternatives to seed-and-extend alignment.
- Fine-grained index partitioning for alignment.

Alignment

Wish list for higher-productivity bioinformatics

- DSL for data cleaning/filtering/wrangling
- Standardized compressed binary file formats
- Programming model support for asynchronous parallel I/O and computation

Thank you!

- Feedback, Questions?

- sites.psu.edu/XPSGenomics
- sprite-psu.sourceforge.net

- Postdoc opening on this project