A Distributed and High-Throughput Short Read Processing Suite

Luca Pireddu*, Simone Leo and Gianluigi Zanetti
CRS4, Polaris, Ed. 1, I-09010 Pula, Italy
*luca.pireddu@crs4.it

Introduction

Seal is a suite of open source tools for *distributed* short read processing designed for high-throughput sequencing operations. It currently provides tools to implement much of the typical sequence processing pipeline, and we aim to release the other necessary tools shortly.

Features

Distributed

Use or rent a cluster and shorten your analysis time.

Scalable

Need to process faster? Simply get more machines! (we've tested Seal with up to 128 nodes)

No central storage

Seal removes the requirement for a high-performance shared storage appliance.

Resistant

Being based on Hadoop, Seal tools can resist node failures and transient cluster conditions like load peaks or network outages.

Who uses it?

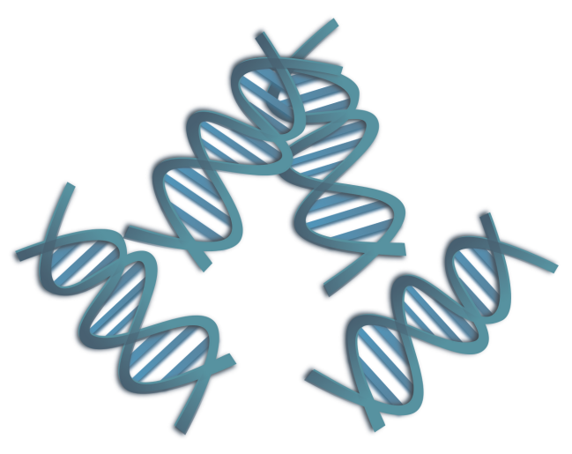
The Seal suite is currently used to implement most of the production pipeline at the CRS4 Sequencing and Genotyping Platform, currently processing data from 6 Illumina sequencing machines. It is also being evaluated for use at the Dutch National Center for Scientific computing.

Libbwa

Seqal, Seal's alignment tool, is based on libbwa, which is a BWA [1] alignment library we built by refactoring the original BWA source code. Libbwa takes pre-loaded data structures as input as opposed to files, and uses mmapped I/O to access the reference index allowing the same reference to be shared in memory by multiple processes. This feature makes it possible to **run in parallel 8 alignments with less than 16 GB of RAM** using the 1KG reference.

Pydoop

Libbwa provides high-level Python bindings for BWA. To create a Hadoop-based alignment tool that could take advantage of this feature we used Pydoop [2], which allowed us to easily run a distributed Python program on the Hadoop framework. In addition, Pydoop provides a Python API for HDFS which we use to easily implement "smart" launcher scripts in Python.



Demux

Distributed utility to demultiplex data from multiplexed Illumina runs.

Seal's Demux replaces the standard Illumina demultiplexer with a Hadoop-based implementation which can drastically reduce run times when run on a cluster and can leverage HDFS storage like all Seal tools.

PairReadsQseq

A distributed utility to convert Illumina qseq files into prq file format: id, read 1, qual 1, read 2, qual 2.

In addition, PairReadsQseq filters read pairs when they don't have a user-defined number of known bases or if failed the machine quality checks.

Seqal

A distributed read mapping (equivalent to BWA aligner) and duplicate removal tool (same criteria as Picard MarkDuplicates).

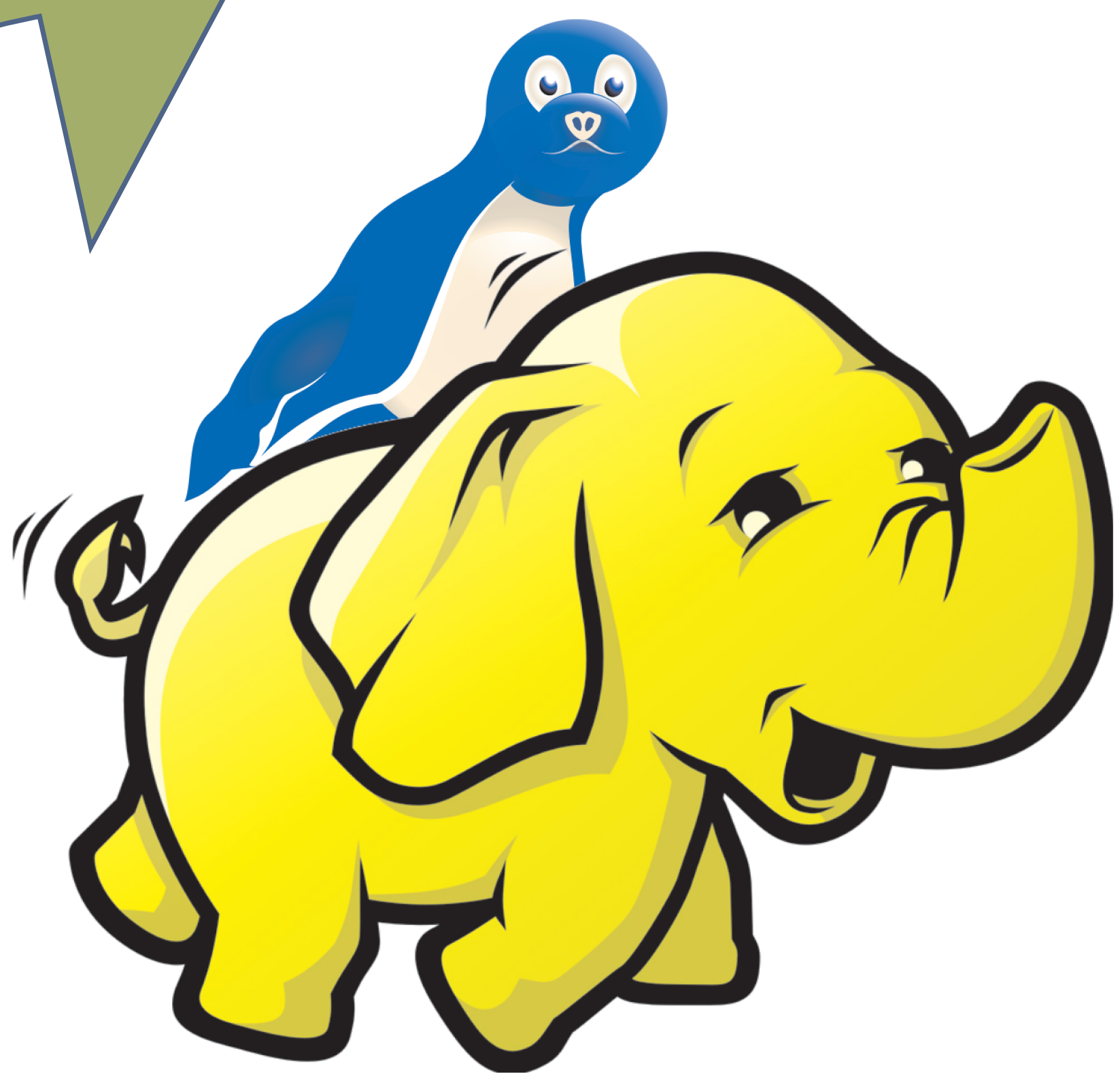
While computing the same paired read alignments as BWA, Seqal can distributed its work over many computing nodes achieving very high throughputs.

ReadSort

A distributed Hadoop utility to sort read alignments by mapping coordinate. ReadSort uses the same sorting algorithm used by Yahoo to win the TeraSort benchmark.

Recalibrate

A distributed program to recalibrate base qualities based on empirical observations for the specific run (like GATK CountCovariates and TableRecalibration). This program is not yet available but is currently under development.



Seal

When are you going to do the laundry?
Haven't you read my genome report? I don't do laundry!
XY
herman roozen
© COMIC HOUSE

Alignment validation

We compared Seal's alignments to BWA ver. 0.5.8c using 5M read pairs from SRA and the UCSC HG18 reference genome: 99.5% *of the alignments were* identical.

The remaining 0.5% had slightly different map quality scores, while the mapping coordinates were identical for all but two reads (which had multiple best hits). All these discrepancies are explained by changes in the insert size statistics, which depend on the specific input read batch.

Scalability/Performance

We tested Seal with varying input size (50 to 400 GB, reads 100 bp long) and cluster size (16 to 96 nodes) and found that it scales well in the number of nodes available. In particular, the figure to the right shows for the alignment and duplicate removal operations (alignment being one of the most CPU-demanding steps the pipeline) that as the number of nodes in increases Seal maintains its throughput per node, as long as the workload is high enough to occupy the machines.

In alignment, Seqal is generally capable of throughputs per node comparable to single-node BWA operation; we measured about 1100 pairs/sec on the same hardware used for the Seal tests. This observation implies that Seqal and Hadoop keep the distribution overhead to a minimum.

Node configuration: dual quad-core Intel Xeon CPUs @ 2.83 GHz, 16 GB RAM, two 250 GB SATA disks (one for HDFS), and GbE.

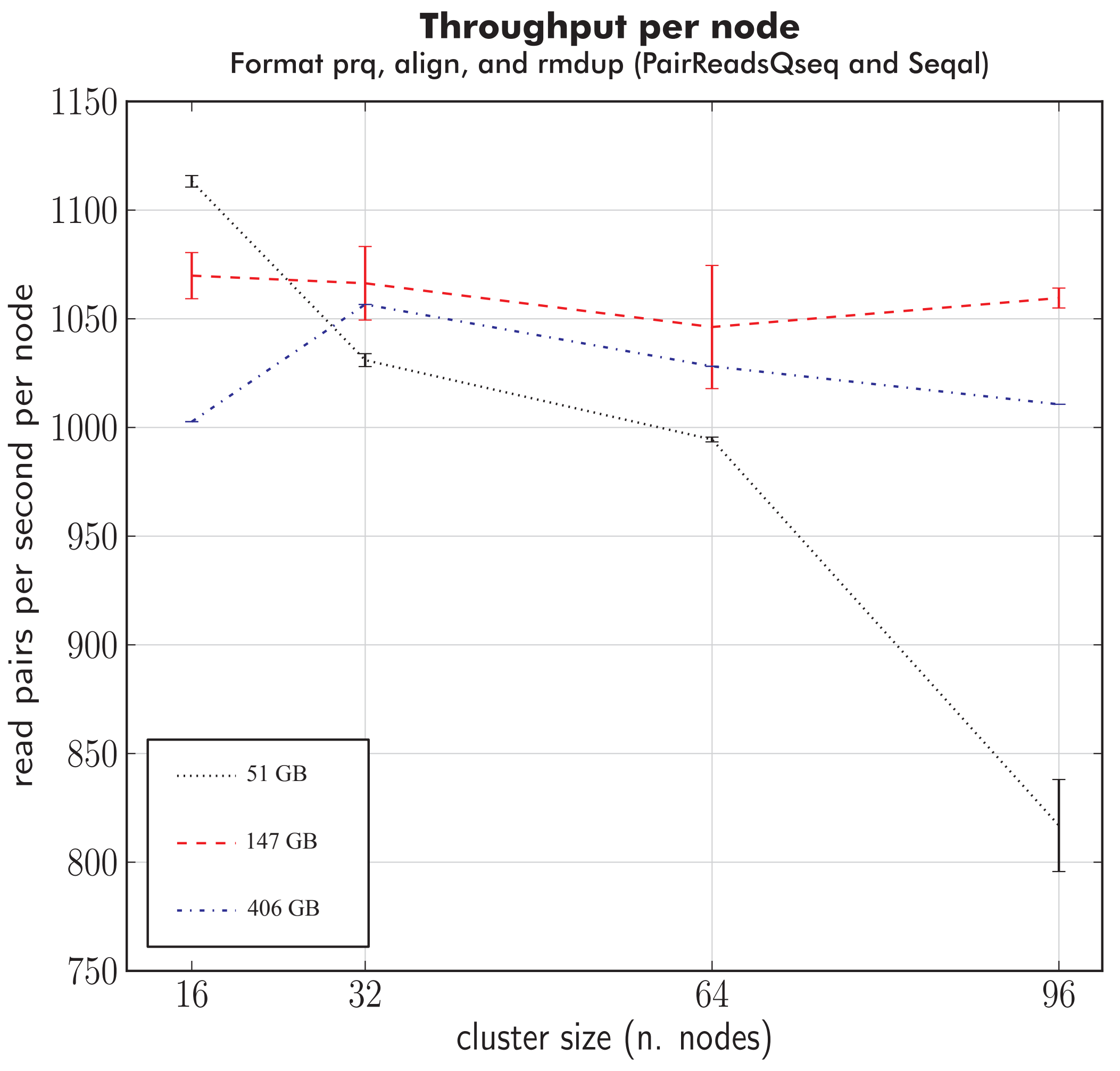


Fig. 1. Throughput per node of a Seal workflow converting qseq to prq, aligning and removing duplicate reads (PRQ and Seqal applications) [3]. An ideal system would produce a flat line. By comparison, a single-node workflow we wrote for testing, using the standard multi-threaded BWA and Picard, reaches approx. 1100 pairs/sec on a 5M pair data set.

BWA alignment and Seal alignment
Runtime comparison

Data set	BWA time (h, 1 node)	Seqal time (h, 32 nodes)
5M	0.49	0.04
1 lane	11.26*	0.63
3 lanes	32.39*	1.72
8 lanes	89.35*	4.78

Table 1. Running times of multithreaded BWA on a node with 8 cores and Seal alignment without duplicates removal on 32 nodes. Seal times include qseq to prq conversion.
*) Linear extrapolation of the throughput observed on the 5M data set.

References

[1] Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25(14), 1754–1760.
[2] Leo, S. and Zanetti, G. (2010). Pydoop: a Python MapReduce and HDFS API for Hadoop. In Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing, pages 819–825.
[3] Pireddu,L., Leo,S. and Zanetti,G. (2011). SEAL: a Distributed Short Read Mapping and Duplicate Removal Tool. Bioinformatics.

http://biodoop-seal.sourceforge.net/