CS-BWAMEM: A fast and scalable read aligner at the cloud scale for whole genome sequencing

Yu-Ting Chen 1,∗, Jason Cong 1, Jie Lei 1, Sen Li 1, Myron Peto 2, Paul Spellman 2, Peng Wei 1, and Peipei Zhou 1

1Computer Science Department, University of California, Los Angeles, CA 90095, USA
2Department of Molecular and Medical Genetics, Oregon Health and Science University, Portland, OR 97239, USA

ABSTRACT
Summary: The deep-coverage whole-genome sequencing (WGS) can generate billions of reads to be sequenced. It is time consuming for state-of-the-art aligners, such as BWA-MEM, to align the tremendous number of reads onto the reference genome. Inherently, the reads can be aligned using a massively parallel approach, and the alignment process should not be bounded by the limited number of computing cores of a single server. We present cloud-scale BWAMEM (CS-BWAMEM), an ultrafast and highly scalable aligner built on top of cloud infrastructures. It leverages the abundant computing resources in a public or private cloud to fully exploit the parallelism obtained from the enormous number of reads. With CS-BWAMEM, the pair-end whole-genome reads (30x) can be aligned within 80 minutes in a 25-node cluster with 300 cores.

Contact: ychen@cs.ucla.edu

1 INTRODUCTION
Aligning reads to the reference genome is the first and one of the most time-consuming steps in many genomics pipelines, such as BWA-MEM, to align the tremendous number of reads onto the reference genome. Inherently, the reads can be aligned using a massively parallel approach, and the alignment process should not be bounded by the limited number of computing cores of a single server. We present cloud-scale BWAMEM (CS-BWAMEM), an ultrafast and highly scalable aligner built on top of cloud infrastructures. It leverages the abundant computing resources in a public or private cloud to fully exploit the parallelism obtained from the enormous number of reads. With CS-BWAMEM, the pair-end whole-genome reads (30x) can be aligned within 80 minutes in a 25-node cluster with 300 cores.

2 METHODS
MapReduce is one of the most frequently used programming models for processing distributed data in a cloud. Apache Spark (Zaharia et al., 2012) implements a novel and efficient MapReduce system that caches intermediate data, which can be reused in future steps, in memory to avoid unnecessary slow disk I/O accesses. The data in a cloud is usually stored in a distributed file system, such as HDFS (Shvachko et al., 2010). In CS-BWAMEM, Spark is used as our computing framework, while HDFS is used to store WGS data.

Fig. 1. Results comparison between CS-BWAMEM and BWA-MEM on (a) runtime, and (b) the number of aligned reads according to the mapping quality scores on a buffy sample.

REFERENCES

* to whom correspondence should be addressed

We have CPUs with different number of cores. The aggregated computing power of our cluster is approximately 25 nodes.