ADAM: Genomics Formats and Processing Patterns for Cloud Scale Computing

Matt Massie[1], Frank Austin Nothaft[1], Chris Hartl[2], Christos Kozanitis[1], and David Patterson[1]

[1]Department of Computer Science, University of California, Berkeley
[2]The Broad Institute of MIT and Harvard

Abstract

Current genomics applications are dominated by the movement of data to and from disk. This data movement pattern is a significant bottleneck that prevents these applications from scaling well to distributed computing clusters. In this report, we introduce a new set of data formats for genomics applications that are designed for in-memory MapReduce processing. These formats improve application performance, data storage efficiency, and programmer productivity.

1 Introduction

The process of transforming reads from alignment to variant-calling ready reads involves several processing stages including duplicate marking, base score quality recalibration, and local realignment. Traditionally, these stages have involved reading a Sequence/Binary Alignment Map (SAM/BAM) file, performing transformations on the data, and writing this data back out to disk as a new SAM/BAM file [1].

References

[1] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., et al. The sequence alignment/map format and SAMtools. *Bioinformatics 25*, 16 (2009), 2078–2079.