Analyzing Large Scale Genotype Datasets With Gnocchi

Frank Austin Nothaft

Abstract

The development of inexpensive DNA sequencing technologies has enabled projects that sequence large cohorts. While many recent projects have tackled the computationally expensive process of turning raw DNA sequence into genomic variants, cohort analyses still rely on traditional single node techniques. To address this problem, we introduce Gnocchi, a Spark SQL based toolkit for analyzing genomic variants. Gnocchi extends the ADAM framework for analyzing genomic data with several variation specific patterns, such as matrix and genotype state views. With Gnocchi, we are able to parallelize common expensive tasks, such as the training of genome-wide assocation models, or large scale population stratification.

1 Introduction

ADAM [1, 2].

References

[1] Massie, M., Nothaft, F., Hartl, C., Kozanitis, C., Schumacher, A., Joseph, A. D., and Patterson, D. A. ADAM: Genomics formats and processing patterns for cloud scale computing. Tech. rep., UCB/EECS-2013-207, EECS Department, University of California, Berkeley, 2013.

[2] Nothaft, F. A., Massie, M., Danford, T., Zhang, Z., Laserson, U., Yeksigian, C., Kottalam, J., Ahuja, A., Hammerbacher, J., Linderman, M., Franklin, M., Joseph, A. D., and Patterson, D. A. Rethinking data-intensive science using scalable analytics systems. In *Proceedings of the International Conference on Management of Data (SIGMOD '15)* (2015), ACM.