Mango: data exploration on large genomic datasets

Alyssa Morrow[1], Eric Tu[2], Frank Austin Nothaft[3], Anthony Joseph[4], David Patterson[5]

[1] University of California-Berkeley, United States. Email: akmorrow@berkeley.edu

[2] University of California-Berkeley, United States. Email: erictu@berkeley.edu

[3] University of California-Berkeley, United States. Email: fnothaft@berkeley.edu

[4] University of California-Berkeley, United States. Email: adj@berkeley.edu

[5] University of California-Berkeley, United States. Email: pattrsn@cs.berkeley.edu

Project Website: http://bdgenomics.org/

Source Code: https://github.com/bigdatagenomics/mango

Current genomics visualization tools are intended for a single node environment and lack the scalability required to visualize multiple whole genome samples. Data from the 1000 Genomes Project provides 1.6 terabytes of variant data and over 14 terabytes of alignment data. However, typical genomic visualizations materialize less than 10 kbp, only 3.3e-7% of the genome. Mango is a visualization browser that selectively materializes and organizes genomic data to provide fast in-memory queries. Mango materializes data from persistent storage as the user requests different regions of the genome. This data is efficiently partitioned and organized in memory using interval trees. This interval based organizational structure supports ad hoc queries, filters, and joins across multiple samples at a time, enabling exploratory interaction with genomic data.

Mango is built on top of Spark and ADAM, both open source projects under the Apache license. Leveraging Spark as Mango's cluster computing framework enables scalable, distributed computations on terabytes of genomic data. Mango leverages ADAM's genomic file formats which can be stored in persistent storage and accessed by Spark. Both ADAM and Mango are part of the Big Data Genomics project at University of California-Berkeley. Mango is published under the Apache 2 license.