## ABSTRACT #4

# OPEN SOURCE TOOLS TO EXPLOIT DNA SEQUENCE DATA FROM LIVESTOCK SPECIES

Derek M. Bickhart[1], Jana L. Hutchison[1], Lingyang Xu[2,3], Jiuzhou Song[3], George E. Liu[2]

[1]USDA, ARS, Animal Improvements Program Laboratory, BARC
[2]USDA, ARS, Bovine Functional Genomics Laboratory, BARC
[3]University of Maryland, Department of Animal and Avian Sciences, College Park, MD

Next-Generation Sequencing (NGS) is a recent technological development that allows researchers to rapidly determine the DNA sequence of an individual. The decrease in cost of NGS has brought the technology into the realm of practical applications in livestock genomics, where it can be used to generate new genetic markers or identify variants that influence productive traits. The amount of data that NGS experiments generate is not trivial (~ 600 gigabytes per flowcell) and it is often the data analysis step that serves as the time bottleneck for sequence based studies. Further confounding the use of this data in agricultural research is the fact that freely-available computational tools designed to process and analyze sequencing data are still limited to specialized programs or species-specific software suites. In order to provide the agricultural community with a tool that can process sequencing data and yield important genetic variants, we have developed a new open source pipeline that incorporates cutting edge software and new algorithms.

Our pipeline consists of a central master process that delegates tasks to sub-processes using a “divide-and-conquer” strategy. This strategy allows the pipeline to make use of as many cpu threads as the user designates, thereby speeding up calculations at a nearly linear rate per processor. The pipeline uses the Broad Institute’s GATK package to call single nucleotide polymorphisms (SNP) and insertions/deletions in DNA sequence. Copy number variations (CNV) are called using custom software designed specifically for this pipeline using the Java programming language. All variant calls are annotated and formatted into an excel-readable spreadsheet, highlighting important information for the user. In order to test the efficiency and utility of our pipeline, we called CNVs and SNPs on a dataset of 72 bulls from eight different cattle breeds sequenced at the USDA’s Bovine Functional Genomics Laboratory.

Starting with a dataset of ~ 4 terabytes of uncompressed sequence data, the pipeline took 26 days to process the results and return SNP and CNV calls for each individual animal using only 20 processor cores and 100 gigabytes of ram. Because of the scalability of the pipeline, we estimate that using an additional 80 processor cores would have reduced total processing time to 5.2 days. An average of 6496 CNVs and ~2 million SNPs were detected in each individual prior to quality filtering and data merger. The pipeline’s annotation tools highlighted several genetic differences among the different breeds, particularly in regions dominated by immune-system related genes. We expect that this pipeline will reduce the time and effort required to analyze NGS data in future sequencing projects and will provide researchers with critical information on their samples without needless frustration.