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

More Information

References

[1] Bickhart, et al. 2012. Copy Number Variation of Individual Cattle Genomes using Next-Generation Sequencing. Genome Research. 22: 778-790
[2] Mils, et al. 2011. Mapping Copy Number Variation by Population-Scale Genome Sequencing. Nature. 470. 59-65

Project Source Code (Alpha Stage)

sourceforge.net/projects/cosvard/

This work was supported by NRI/AFRI grant no. 2011-67015-30183 from the USDA NIFA

Contact information

Derek Bickhart
USDA ARS AIPL

derek.bickhart@ars.usda.gov

Phone: (301) 504 - 8592

The Problem: DNA Sequence data allows researchers to identify variations within the genomes of individuals of a species that could impact important production traits; however, tools designed to identify these variations are often designed specifically for mouse or human studies.

Proposed Solution: We are designing an open-source pipeline for the analysis of DNA sequencing studies involving non-model organisms and agricultural species. The pipeline will be free for academic and research use and the source code for all written programs will be released with the final package. Additionally, the results of the analysis will be provided to the end-user in easily accessible forms, such as excel spreadsheets, text summaries and pdf file plots.

(4) Variant calling

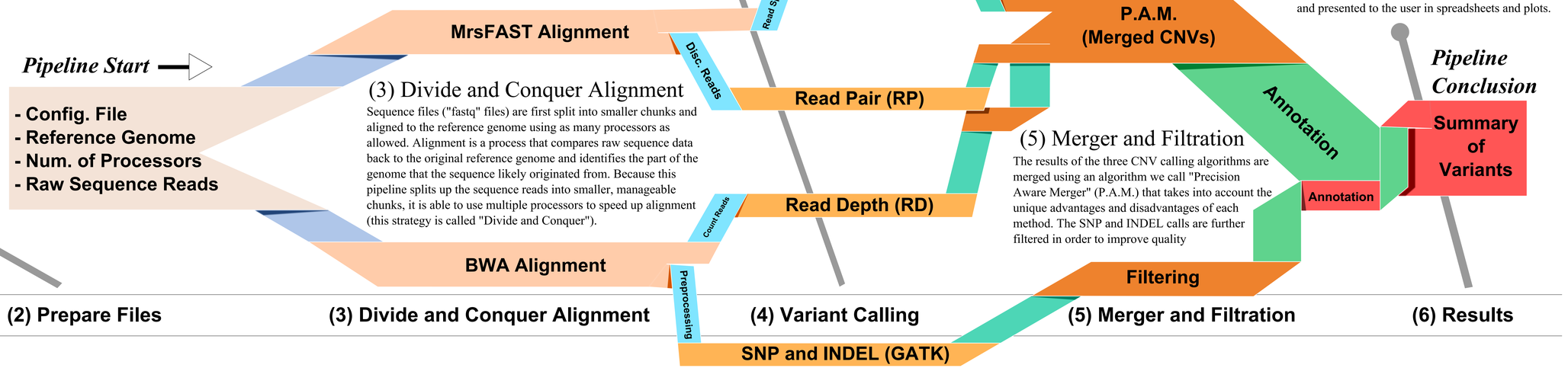
Aligned sequence data is then processsed using several programs in order to identify Copy Number Variants (CNVs), Single Nucleotide Polymorphisms (SNPs) and Insertion/Deletions (INDELs). CNVs are called using three algorithms used by the Human 1000 Genomes Project [2] (SR, RP and RD) whereas SNPs and INDELs are called by the Broad Institute's Genome Analysis Toolkit (GATK).

(6) Final Results

CNV, SNP and INDEL calls are then cross referenced against information known about the genome in order to see if they impact genes or other functional genetic regions (this is called "Annotation"). Afterwards, the results are tabulated and presented to the user in spreadsheets and plots.

(2) Prepare Files

In order to make management of each step more flexible, we have implemented a configuration file control system for running the pipeline. The configuration file contains the locations of sequence files ("fastq" files) and the locations of installed pipeline programs and scripts. In order to speed up computation time, the pipeline can make use of multiple processor cores and high performance computing architectures.



(3) Divide and Conquer Alignment

Sequence files ("fastq" files) are first split into smaller chunks and aligned to the reference genome using as many processors as allowed. Alignment is a process that compares raw sequence data back to the original reference genome and identifies the part of the genome that the sequence likely originated from. Because this pipeline splits up the sequence reads into smaller, manageable chunks, it is able to use multiple processors to speed up alignment (this strategy is called "Divide and Conquer").

(5) Merger and Filtration

The results of the three CNV calling algorithms are merged using an algorithm we call "Precision Aware Merger" (P.A.M.) that takes into account the unique advantages and disadvantages of each method. The SNP and INDEL calls are further filtered in order to improve quality

(1) DNA Resequencing

The Human Genome Project, which was the first major DNA sequencing project for a large Eukaryotic genome, lasted 10 years and cost approximately $3 billion. The cost to sequence new genomes has dropped precipitously since that project concluded, and current price estimates hover around $6,000 to $8,000 per individual sequenced. Much of this price reduction is due to new methods of sequencing and improved instrumentation (e.g. the HiSeq 2000; pictured right inset).

Our pipeline was designed for the explicit purpose of analyzing DNA sequence data from organisms that already have a finished reference genome project. As of the writing of this poster, this includes 6 agricultural animal species and 17 agricultural plant species with many more to shortly follow. The reason why researchers resequence individuals of a species that already have a completed reference genome is to identify variations in DNA sequence within that individual's genome. These variations can be linked to disease susceptibility (powdery mildew susceptibility in Arabidopsis (a)) or productive traits (white coat color in Sheep (b)).

A Test Run: 100 Sequenced Bulls

Cattle Breed	Number of Animals	Millions of base pairs (Mbp)
Bos indicus Brahman	7	116.98
Bos indicus Gir	6	146.87
Bos indicus Nelore	8	170.44
Bos taurus Angus	17	1027.09
Bos taurus Holstein	32	718.95
Bos taurus Jersey	8	144.6
Bos taurus Limousin	7	155.47
Bos taurus Romagnola	4	92.57

Table: Our dataset was composed of eight different breeds of cattle from two cattle subspecies: *Bos indicus* (or "zebu") and *Bos taurus*. Each individual was sequenced to at least 6X coverage, and several animals were sequenced to greater depth in order to provide a contrast.

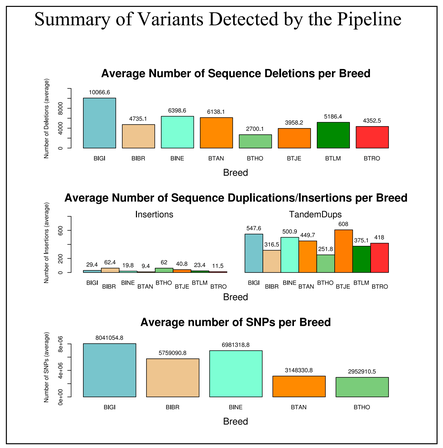


Figure: A summary of all variants currently detected using our pipeline shows a clear difference between indicus and taurus subspecies in the number of SNPs identified. Insertions and deletions vary far more among breeds than subspecies and may reflect smaller phenotypic differences.

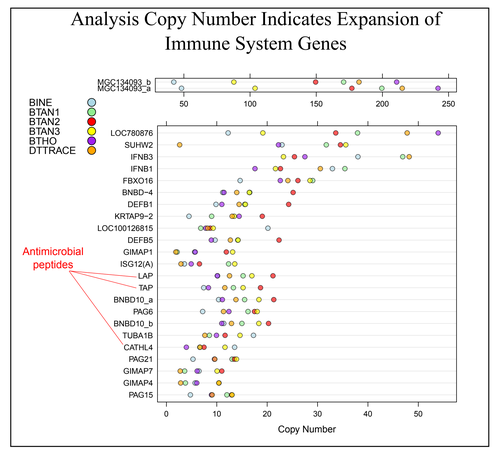


Figure: Association of genetic features with variants (Annotation) has revealed several interesting biological features [1]. Antimicrobial peptides, which serve as a first line of defense for the immune system, are often duplicated. This may indicate an evolutionary "arms-race" between the animal and environment as bacteria evolve to resist Antimicrobial compounds over time.