Analysis of Natural Variation in 30 Sorghum Landraces

Riley Ziegler[1] and Charles Addo-Quaye[1]

[1]Division of Business and Computer Science, The Lewis-Clark State College, Lewiston, ID 83501

Abstract

Sorghum is a next generation of crop species for food grain, feedstock, beverage and biofuel production. To discover highly desirable agronomic traits in sorghum, we analyzed 3.42 billion DNA sequences derived from 30 sequenced sorghum landraces using next-generation sequencing (NGS) technology. Using the BWA short reads aligner, 97% of the sequenced reads mapped successfully to the sorghum reference genome. Using the SAMtools variant-calling algorithm, we detected 68.14 million mutations, including 61.32 million DNA base substitutions or single nucleotide polymorphisms (SNPs) and 6.81 million insertions and deletions (INDELs). In our preliminary analysis using the snpEff variant annotation tool, we predicted a total of 134,207 high-impact mutations and 1.81 million moderate-impact mutations in the 30 sequenced sorghum landraces.

Figure 1. Field of sorghum bicolor maturing for harvest
Source: https://www.southernliving.com/what-is-sorghum-7094287

Methods

We retrieved next-generation sequencing (NGS) data for 30 sorghum landraces (Figure 1). Figure 2 shows our computational pipeline for mutation prediction. Utilizing the *BWA* short reads aligner we indexed and mapped the NGS reads to the sorghum reference genome version 3.1. We were able to get a binary alignment map (BAM) for use in the variant calling step. Using *SAMtools* & *BCFtools* to process the BAM files, we created Variant Call Files (VCF) containing the predicted mutations in each sorghum individual. After building a database for sorghum using the *snpEff* variant annotation tool, we annotated each VCF file to classify which mutations were low, moderate, and high-impact on the gene function in the sorghum individuals.

Table 1. Whole-genome NGS sequencing statistics for the sorghum population.

	Results
Total Number of Reads	3,424,760,730
All Mapped Reads	3,351,467,216
Properly Paired Mapped Reads	3,056,922,402
Average Number of Reads	114,158,691

Results

Table 1 shows, that out of 3.42 billion reads, 98% mapped to the reference genome and 89% were properly paired. After filtering and sorting the alignments we found a total of 68.14 million natural variants (Table 2), including 61.32 million single nucleotide polymorphisms (SNPs) and 6.81 million insertions and deletions (INDELs). Figure 2 shows the distribution of predicted mutation qualities, and shows a majority were of high quality (225). Table 3 shows the *snpEff* classification of the mutation impacts on the 30 sorghum landraces; 134,207 are high-impact, 1.81 million moderate, and 1.88 million low. Of the 1.88 million low-impact mutations, 1.5 million are synonymous amino acid changes.

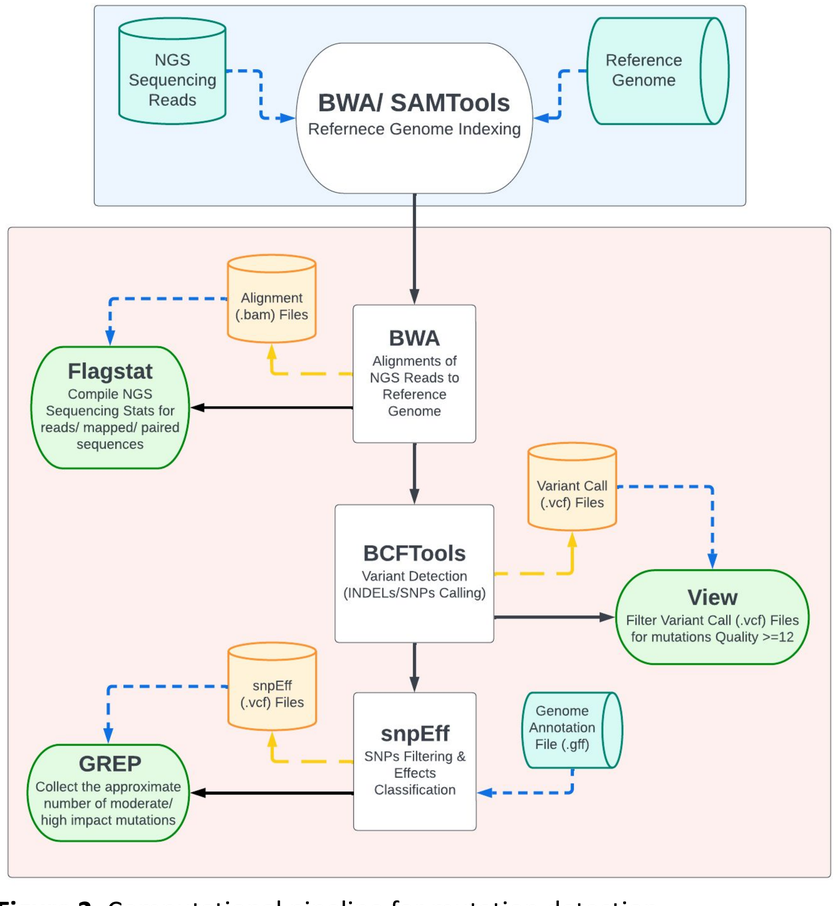

Figure 2. Computational pipeline for mutation detection.

Table 2. Results of mutation detection in the 30 sorghum landraces

Variants	Total
All Variants	68,141,819
INDELs	6,813,482
SNPs	61,328,337
INDELs (Qual > 20)	6,813,482
SNPs (Qual > 20)	61,328,337

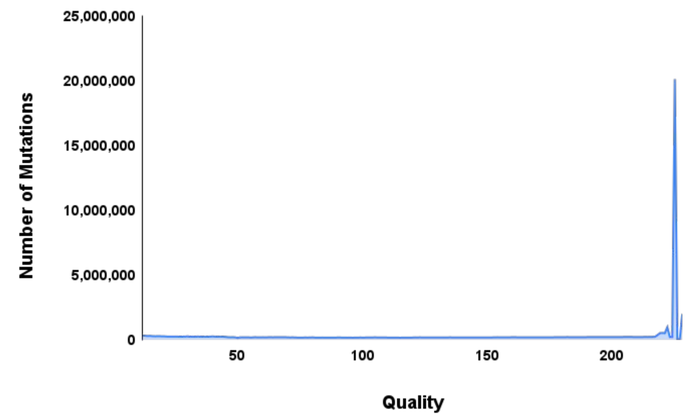

Figure 2. Distribution of mutation qualities from SAMTools variant-calling

Table 3. Distribution of the different impact levels for a given mutation

SNP Impact	Total	Average
High Impact	**134,207**	**4,473.57**
Stop Gained	319,743	62,722.60
Stop Lost	17,918	60,425.60
Start Lost	32,075	2,305,891.75
Splice Site Donor	10,676	17,061.43
Splice Site Acceptor	64,462	69,460.03
Moderate Impact	**1,812,768**	**2,148.73**
Missense	2,083,801	355.87
Low Impact	**1,881,678**	**1,069.17**
Synonymous	1,505,792	597.27
Start Gained	511,843	10,658.10
Modifier	**69,176,749**	**50,193.07**

References:

1. Li, Heng, and Richard Durbin. "Fast and accurate short read alignment with Burrows–Wheeler transform." bioinformatics 25.14 (2009): 1754-1760.
2. Li, Heng, et al. "The sequence alignment/map format and SAMtools." bioinformatics 25.16 (2009): 2078-2079.
3. Cingolani, Pablo, et al. "A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3." fly 6.2 (2012): 80-92.

Acknowledgement

This research is supported by a grant from the Idaho State Board of Education's Higher Education Research Council (HERC).