In silico detection of EMS-induced mutations in an Arabis alpina population.

Beknazar Tursyngazy[1] and Charles Addo-Quaye[1]

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

Abstract

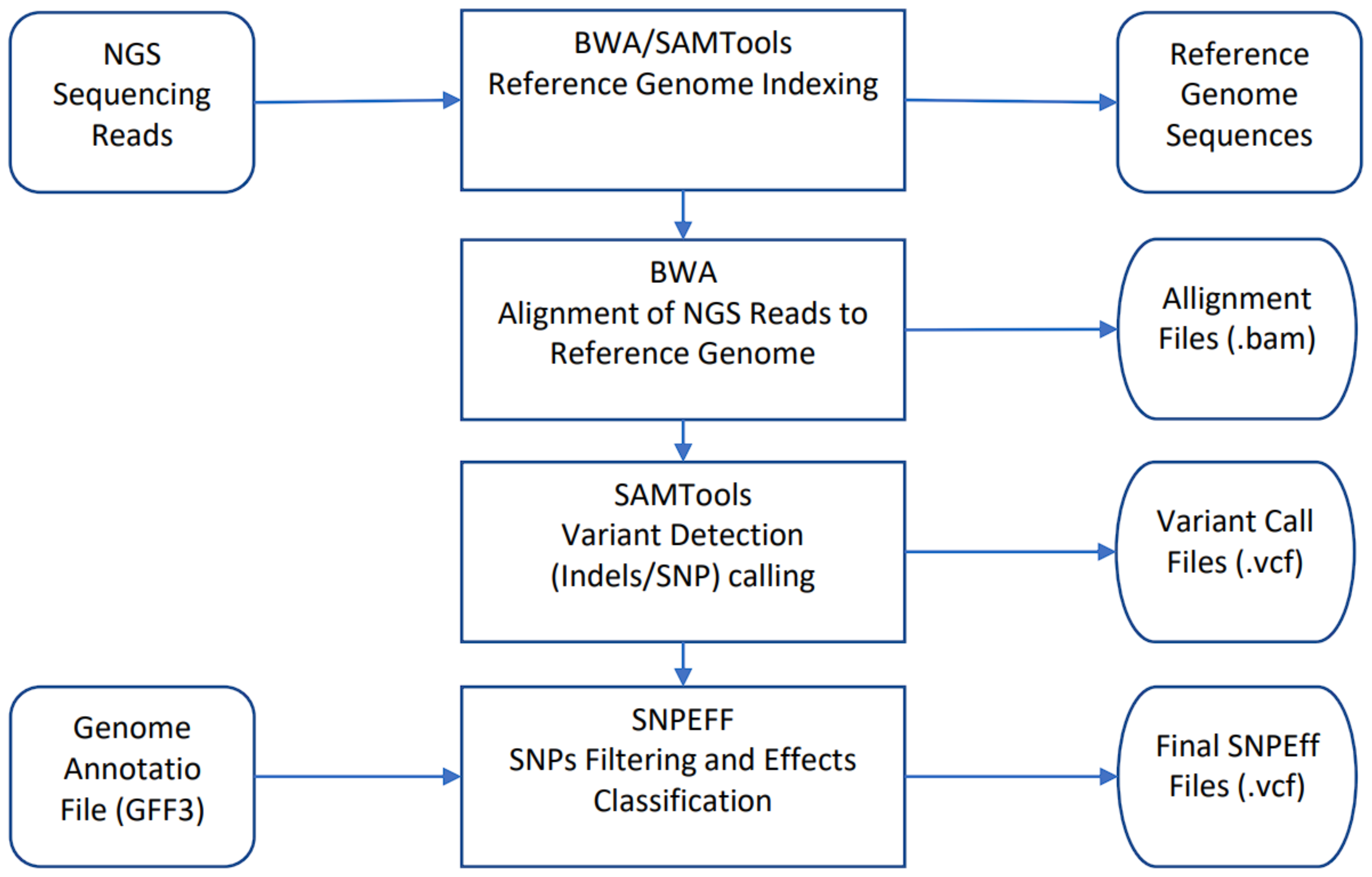
Arabis alpina (Alpine rock-cress weed) is a flowering plant, native to mountainous environments of the northern hemisphere. We analyzed 1,454,931,853 next-generation sequencing (NGS) reads from 38 sequenced Arabis alpina mutant individuals which that were mutagenized using the chemical mutagen, ethyl methane sulfphonate (EMS). Using the BWA short reads mapper, BWA, 95% (1,387,167,658) of the NGS reads mapped to Arabis alpina reference genome version 4. Using the SAMtools variant- detection algorithm, SAMtools, we detected a total of 1,457,917 mutations, with an average of 38,366 mutations per sample. Overall, the predicted mutations include 971,252 high-quality single nucleotide polymorphisms (SNPs) and 168,783 high-quality insertions and deletions (INDELs).

Methods

We obtained NGS whole-genome sequencing data for 38 *Arabis alpina* (Figure 1) EMS-mutagenized individuals from the NCBI Sequenced Reads Archive (SRA) repository. Figure 2 shows our implemented DNA mutation prediction pipeline. Next, after indexing the reference genome (version 4.0) of *Arabis alpina* by using BWA and SAMtools, we mapped the pre-processed short reads from the mutant genomes to the reference genome by using the BWA short reads aligner. After aligning the reads to the reference genome, SAMtools was used to call the variants which could be single nucleotide polymorphisms (SNPs) or small insertions or deletions (INDELs) from the alignment file. Finally, to predict the effect or impact of the mutations on gene function, we used the snpEff software to annotate the mutations and create the variant call files (VCFs). The annotated VCF file provides valuable information about the variants present in the genomes and their functional impact.

Figure 2. Image of Arabis alpina (Alpine rock-cress weed).

Source: https://depositphotos.com/328529872/stock-photo-caucasian-rock-cress-flower-garden.html



NGS Sequencing Reads

BWA/SAMTools Reference Genome Indexing

Reference Genome Sequences

BWA Alignment of NGS Reads to Reference Genome

Allignment Files (.bam)

SAMTools Variant Detection (Indels/SNP) calling

Variant Call Files (.vcf)

Genome Annotatio File (GFF3)

SNPEFF SNPs Filtering and Effects Classification

Final SNPEff Files (.vcf)

Figure 1. Pseudocode describing the *Arabis alpina* annotation algorithm.

Results

We obtained 1,454,931,853 NGS reads of which 1,387,167,658 mapped to the *Arabis alpina* reference genome, with an average of 38,287,680 reads per mutant genome (Table 1). Overall, we predicted 1,457,917 EMS-induced mutations including 222,578 INDELs and 1,235,339 SNPs (Table 2). Annotation of mutations revealed 17,830 high impact mutations and 50,153,551 moderate impact mutations including an average of 469 high impact mutations per mutant (Table 3).

Table 1. Number of reads after using bwa software.

	Counts
Total Number of Reads	1,454,931,853
All Mapped Reads	1,387,167,658
Paired Mapped Reads	1,443,273,406
Average Number of Reads	38,287,680

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).

Table 2. The nearly 1,4 million variants of mutation were detected, more than 1,1 million of them are high quality.

Variants	Total
All Variants	1,457,917
INDELs	222,578
SNPs	1,235,339
Q20 Variants	1,140,035
Q20 INDELs	168,783
Q20 SNPs	971,252

Table 3. From total of 38 *Arabis alpina individuals we detected 17,830 high impact mutations with an average of 469 and 50,153,551 moderate impact mutations with an average of 1,319,830.*

SNP Impact	Total	Mean
High Impact	17,830	469
Stop Gained	232	6
Stop Lost	3,120	82
Splice Site Donor	2,688	71
Splice Site Acceptor	452	12
Moderate Impact	50,153,551	1,319,830
Missense	162,199	4,268
Low Impact	26,637	701
Silent	26,637	701
Start Gained	19,889	523
Modifier	165,276	4,349