GBCF Bioinformatics Analysis Report

Project: Chestnut EST SSR Analysis
Contact: Jeanne Romero-Severson
Analyst: Elizabeth Brooks
Date: 19 January 2023

Analysis Workflows

Prep

For each run:

1. QC (FastQC)
2. Trimming (Trimmomatic)
3. Mapping (BWA)

SSR Lengths

For each run:

1. Measuring SSR lengths (GapGenes.v3.py)
2. Filtering the matrix of SSR lengths (SnipMatrix.py)
3. Formatting of SSR lengths matrix for JoinMap (Format_Matrix.py)

SNP Calling

For each run:

1. Retrieval of sequences flanking SSRs at least 50bp in both directions (SamIAm.py)
2. Sorting and removal of pcr duplicates (SAMtools)
3. Filtering of sequences to keep only unique read alignments (SAMtools)
4. Clipping to soft mask primer sequences (BAMClipper)
5. Addition of sample read groups (SAMtools)

Across all runs:

6. Variant calling (BCFtools)
7. Variant filtering (BCFtools)
8. Variant trimming to remove SSR regions (BEDTools)
9. Formatting of variants matrix (variantMatrix_bcftools.sh)

Notes

- The basic SSR lengths workflow is performed on a per-run basis
- The SNP calling workflow is performed with the samples from all of the available runs (run1 to run8)
- The run number for each sample is appended to the end of each sample name (e.g., SAMPLE1_run1) in the results matrix from the SNP calling workflow

Methods

Raw sequences were trimmed of adapters with Trimmomatic version 0.39 (Bolger et al., 2014) and assessed for quality with FastQC v0.11.8 (Andrews, 2010). Trimmed sequences were aligned to the reference marker contigs using the BWA software package version 0.7.17-r1188 (Li et al., 2009). Corresponding alignments were sorted and sample read groups applied with SAMtools and BCFtools versions 1.9 (Danecek et al., 2021). Primer sequences were soft masked using BAMClipper v1.0.0 (Au et a., 2017). Variants were called and filtered using BCFtools, then trimmed of SSR regions using BEDTools v2.30.0 (Quinlan & Hall 2010). Custom Python2 scripts created previously by Joseph Sarro were used to measure and filter SSR lengths, then create the formatted matrix of SSR lengths for the basic workflow. A custom BASH script was created by Elizabeth Brooks to format the matrix of variants resulting from the SNP calling workflow.

References

Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data[Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for IlluminaSequence Data. Bioinformatics, btu170

Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60. [PMID: 19451168]

Twelve years of SAMtools and BCFtools Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin OPollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, Heng Li GigaScience, Volume 10, Issue 2, February 2021, giab008,https://doi.org/10.1093/gigascience/giab008

Au, C., Ho, D., Kwong, A. et al. BAMClipper: removing primers from alignments to minimize false-negative mutations in amplicon next-generation sequencing. Sci Rep 7, 1567 (2017). https://doi.org/10.1038/s41598-017-01703-6

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics (Oxford, England), 26(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033