Upload image

Version 1

Jul 30, 2020

SNP Calling and VCF Filtering Pipeline V.1

M Renee Bellinger[1]

[1]University of Hawaii at Hilo

1 Works for me

dx.doi.org/10.17504/protocols.io.84fhytn

M Renee Bellinger
University of Hawaii at Hilo

ABSTRACT

SNP-calling and genotype filtering using bowtie2, samtools, bcftools, and vcftools

This pipeline was used for calling SNPs using GBS data obtained from Taro Leaf Blight resistant mapping population of taro, *Colocasia Esculenta* mapped to a taro genome assembled from 10x Genomics linked-reads and Oxford Nanopore Technology long-reads.

The genotype filtering is meant to be done in a single job submission on a linux machine (this was run using 20 cores with 120 Gb RAM and takes two to three days).

The file "1_SNP_calling.txt" contains steps for calling SNPs from demultiplexed GBS data and filtering vcf files (these are broken out in "Steps".

The file "2_genome_analysis" contains steps for running NLR-Annotator and Mummer with Nucmer algorithm to generate synteny plots. This is not broken out in steps

THIS PROTOCOL ACCOMPANIES THE FOLLOWING PUBLICATION

Bellinger, MR, R Paudel, S Starnes, L Kambic, M Kantar, T Wolfgruber, K Lamour, S Geib, S Sim, S Miyasaka, M Helmkampf, M Shintaku. Taro genome assembly and linkage map reveal QTLs for resistance to Taro Leaf Blight. Submitted GigaScience.

ATTACHMENTS

TaroUH_key.txt.csv.C2YWBACXX_Hawaiian.samp_bar_enz.tab
TaroUH_key.txt.csv.C3008ACXX_P230xP255.samp_bar_enz.tab
Taro_UH_report.pdf
keep_1025_230_255.txt
1_SNP_calling.txt
2_genome_analysis.txt

DOI

dx.doi.org/10.17504/protocols.io.84fhytn

PROTOCOL CITATION

M Renee Bellinger 2020. SNP Calling and VCF Filtering Pipeline. **protocols.io**
dx.doi.org/10.17504/protocols.io.84fhytn

MANUSCRIPT CITATION please remember to cite the following publication along with this protocol

Bellinger, MR, R Paudel, S Starnes, L Kambic, M Kantar, T Wolfgruber, K Lamour, S Geib, S Sim, S Miyasaka, M Helmkampf, M Shintaku. Taro genome assembly and linkage map reveal QTLs for resistance to Taro Leaf Blight. Submitted GigaScience.

KEYWORDS

SNPs, GBS, mapping population, vcf, vcf filtering

CREATED

Nov 06, 2019

LAST MODIFIED

Jul 30, 2020

PROTOCOL INTEGER ID

29543

BEFORE STARTING

##############
SNP calling
##############

To set up your working environment you will need to download taro data from NCBI and set paths to each program if they are not already in your bash environment.

The SRA files will need to be demultiplexed

The DNA was digested with enzyme PstI; see file Taro_UH_report.pdf for details

The index keys files are:
TaroUH_key.txt.csv.C2YWBACXX_Hawaiian.samp_bar_enz.tab
TaroUH_key.txt.csv.C3008ACXX_P230xP255.samp_bar_enz.tab # this is the 1025 mapping population

#----------------------------------#
Programs and versions used in pipeline
#----------------------------------#
bowtie2v2.2.4
samtools1.4.1
bcftools-1.2
vcftools-v0.1.14

#----------------------------------#
DATA FILES
#----------------------------------#

GENOTYPING INPUT FILES
Taro GBS data are available from NCBI's Short Read Archive

(1)
SRX2754311 --> GBS data for Taro from Hawaii, South Pacific, Palau, and mainland Asia
the original flow_cell is C2YWBACXX_6_fastq.gz

(2)
SRX2754311 --> GBS data for the '1025' Taro mapping population resistant to Taro Leaf Blight
the original flow_cell is C3008ACXX_8_fastq.gz

GENOME FILE
Download Taro genome from NCBI
Bioproject PRJNA567267

1 ##############

Citation: M Renee Bellinger (07/30/2020). SNP Calling and VCF Filtering Pipeline. https://dx.doi.org/10.17504/protocols.io.84fhytn

```
# SNP calling #
##############

# To set up your working environment you will need to download GBS and genome data from NCBI
# Set paths to each program if they are not in your bash environment.

#---------------------------------#
# Programs and versions used for SNP calling
#---------------------------------#
# bowtie2v2.2.4
# samtools1.4.1
# bcftools-1.2

#---------------------------------#
# DATA FILES
#---------------------------------#

# Short-read files that have been demultiplexed (see Guidelines)
# SRX2754311 --> GBS data for Taro from Hawaii, South Pacific, Palau, and mainland Asia
# SRX2754311 --> GBS data for the '1025' Taro mapping population resistant to Taro Leaf Blight

# GENOME FILE
# NCBI Bioproject PRJNA567267
```

CALL SNPS

2 # copy the genome to a working folder and unzip it to be a "working copy"

```
cp /your/path/to/genome/genome.gz genome.fasta.gz
gunzip genome.fasta.gz
```

3 # index genome

```
samtools faidx genome.fasta
bowtie2-build genome.fasta genome
```

4 # Map demultiplexed GBS reads with bowtie2 or a program of your choice. Repeat this step for all demultiplexed samples
settings are -q for fastq format, -p for 20 threads, --very-sensitive-local for alignment algorithm, -x for reference, -U name of demultiplexed read file (e.g., P230.R1.fastq). -S is output file "P230.sam"

```
bowtie2 -q -p 20 --very-sensitive-local -x genome -U P230.R1.fastq -S P230.sam
```

5 # Use samtools to convert *sam to *bam format, sort the reads, filter out unmapped reads (-F 4) to save space, and then check mapping stats
this loop will process all *sam files in your working directory
note some samtools programs use -o while others use > to collect program output

```
for i in *sam;
do
    samtools view -bS -o $i.bam $i --threads 20
    samtools sort $i.bam -o $i.sort.bam --threads 20
    samtools view -h -F 4 $i.sort.bam > $i.mapped.bam
    samtools flagstat $i.mapped.bam > $i.mapped.sort.stats
done;


#calculate genotype likelihoods and call variants
```

6 samtools mpileup -gu -f genome.fasta -q 20 -Q 20 --output-tags DP,AD,ADF,ADR,SP,INFO/AD *mapped.bam | bcftools call -m --output-type v -o draft_Taro.merged_q20Q20.vcf

7 # compress the *vcf to save space
gzip draft_Taro.merged_q20Q20.vcf

Filter SNPs in *vcf file

8 ## I use variables for this script. Would be easier to run the file 1_SNP_calling.txt provided on page with Abstract

This step remove samples that are not part of the mapping population

A=draft_Taro.merged_q20Q20.vcf.gz

vcftools --gzvcf $A --keep keep_1025_230_255.txt --recode --recode-INFO-all --out 1025_230_255

9 ## Remove SNPs within 5 bases of an INDEL

B=1025_230_255.recode.vcf

bcftools filter -g 5 $B >filter.1.vcf

10 ## Remove indels

C=filter.1.vcf

vcftools --vcf $C --keep keep_1025_230_255 --remove-indels --recode --recode-INFO-all --out 1025_230_255.noindel

11 ## light filter for max missing 50%, minimum depth of 5 (per genotype), bi-allelic calls, minimum quality phred 20
remove samples that are not the 1025 taro leaf blight mapping population

D=1025_230_255.noindel.recode.vcf

vcftools --vcf $D --max-missing 0.5 --minDP 5 --min-alleles 2 --max-alleles 2 --minQ 20 --recode --recode-INFO-all --out 1025_230_255.noindel.maxmiss.5.minDP5.2alleles.minQ20

12 ## Measure missingness across individuals and create a file listing individuals missing >30% of data

E=1025_230_255.noindel.maxmiss.5.minDP5.2alleles.minQ20.recode.vcf

vcftools --vcf $E --missing-indv --out 30percent

cat 30percent.imiss | awk '{if($5>0.3)print $1}'|grep -v INDV> remove-indv_30perc

13 ## Remove individuals missing 30% data in conjunction with the light filter

D=1025_230_255.noindel.recode.vcf

vcftools --vcf $D --remove remove-indv_30perc --max-missing 0.5 --minDP 5 --min-alleles 2 --max-alleles 2 --minQ 20 --recode --recode-INFO-all --out 1025_230_255.noindel.maxmiss.5.minDP5.2alleles.minQ20.n86ind

14 ## Measure missingness across loci and create a file "exclude_miss20p". Apply this file in step 15.

F=1025_230_255.noindel.maxmiss.5.minDP5.2alleles.minQ20.n86ind.recode.vcf

vcftools --vcf $F --keep keep_1025_230_255 --missing-site --out

1025_230_255.noindel.maxmiss.5.minDP5.2alleles.minQ20

cat 1025_230_255.noindel.maxmiss.5.minDP5.2alleles.minQ20.lmiss|awk '{if($6>=0.20)print $1, $2}'>exclude_miss20p

15 ## Apply stringent filter:
Exclude individuals missing 30% of data and loci that are missing 20% of data
some filters are redundant in case steps are run out of order
apply minimum genotypic depth of 8

G=1025_230_255.noindel.recode.vcf

vcftools --vcf $G --remove remove-indv_30perc --exclude-positions exclude_miss20p --max-missing 0.8 --minDP 8 --min-alleles 2 --max-alleles 2 --minQ 20 --recode --recode-INFO-all --out 1025_230_255.noindel.maxmiss.8.minDP8.2alleles.minQ20.n86

16 ## Apply minor allele frequency filter

vcftools --vcf 1025_230_255.noindel.maxmiss.8.minDP8.2alleles.minQ20.n86_variable_parental.recode.vcf.recode.vcf --maf 0.012 --recode --recode-INFO-all --out 1025_230_255.noindel.maxmiss.8.minDP8.2alleles.minQ20.n86_variable_parents_maf_0.012

filtered file is:
1025_230_255.noindel.maxmiss.8.minDP8.2alleles.minQ20.n86_variable_parents_maf_0.012.recode.vcf

Citation: M Renee Bellinger (07/30/2020). SNP Calling and VCF Filtering Pipeline. https://dx.doi.org/10.17504/protocols.io.84fhytn