Nova Lifetech

Your partner in microbiome science

1st BASE

YOUR RESEARCH, OUR BUSINESS

VARIANT CALLING REPORT (BACTERIA)

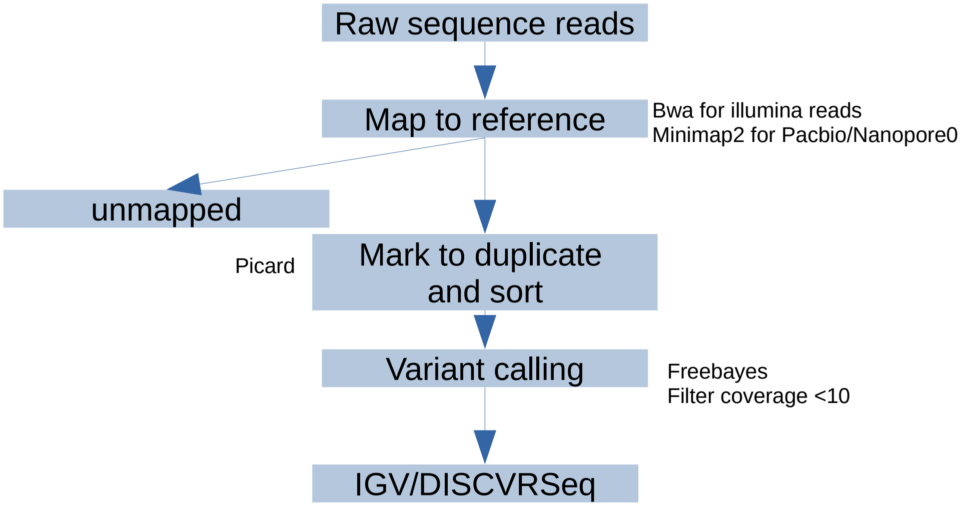


Figure 1. Variant calling workflow using freebayes V1.3.6 (1)

2.1 Workflow

2.1.1 Pre-processing of NGS Raw Reads

For Illumina reads, paired-end reads were first removed of low quality reads (below Phred score Q20) and sequence adaptor using Cutadapt version: 1.18 (2) mplemented in trim-galore V0.650 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). Quality of clean reads were inspected using FastQC (version 0.11.8) and compiled into a single report using MultiQC V1.12 .

2.1.2 Variant calling

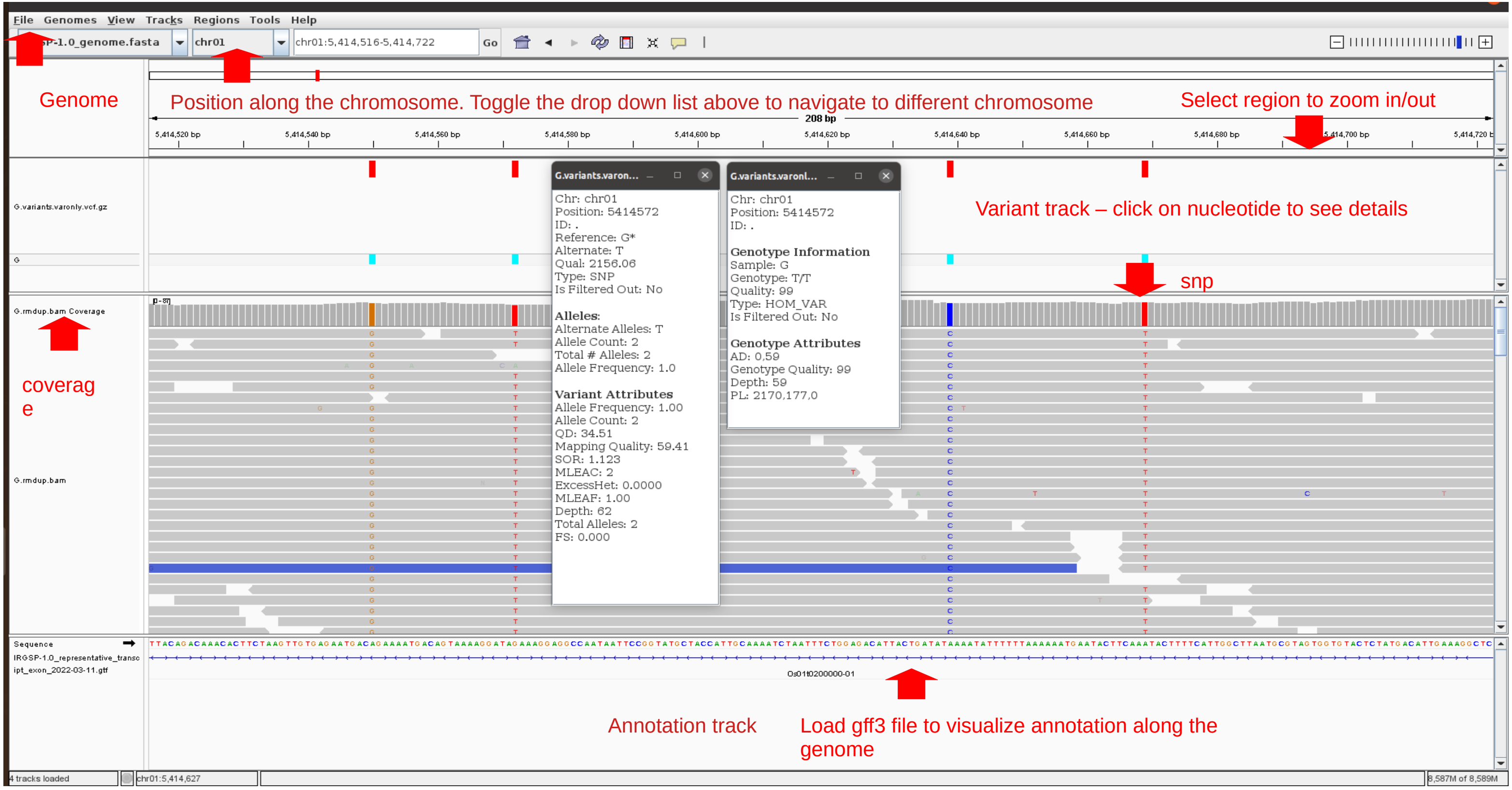
For Illumina paired ends reads, bwa v0.7.17 (3) is used to make index of the reference genome, and mapping of cleaned reads. For long reads such as PacBio’s, mapping to the reference genomeis is done using Minimap2 (4) with the option -ayYL --MD –eqx. Duplicate reads as a results of biases such as those introduced during data generation steps (*e.g.*,PRC amplification) are marked using the *MarkDuplicates* function in Picard and sorted into coordinate-order using the *SortSam* function (https://broadinstitute.github.io/picard/). Variant calling is done using Freebayes V1.3.6 (1). Results is plotted using VariantQC (5).

2.1.3 Visualization of variants

Variants in VCF format can be visualized using the Integrative Genomics Viewer (IGV), a high-performance visualization tool for interactive exploration of large, integrated genomic datasets.

1) Go to https://software.broadinstitute.org/software/igv/download to download and install IGV browser.
2) Choose the correct version of IGV to install on your machine (Linux, Windows, Mac etc)
3) After installation, load the reference genome in fasta format (click on “Genomes > Load Genomes from File….)
4) Additional tracks to be loaded:
   a) Filtered variant calling file, *.variants.varonly.vcf.gz (click on File > Load from File...)
   b) Mapping files after duplicate reads are removed; *.rmdup.bam (click on File > Load from File…)
      This will load the coverage plot on the IGV browser
   c) Annotation file in gtf format (click on File > Load from File…)
      This will load the annotation track and help in identifying genes for which snp is detected
5) Zoom into appropriate section of the chromosome for better clarity

Please see details here: *https://software.broadinstitute.org/software/igv/book/export/html/6*



Genome

Position along the chromosome. Toggle the drop down list above to navigate to different chromosome

Select region to zoom in/out

Variant track – click on nucleotide to see details

snp

coverage

Annotation track

Load gff3 file to visualize annotation along the genome

G.variants.varon...

Chr: chr01
Position: 5414572
ID: .
Reference: G*
Alternate: T
Qual: 2156.06
Type: SNP
Is Filtered Out: No

Alleles:
Alternate Alleles: T
Allele Count: 2
Total # Alleles: 2
Allele Frequency: 1.0

Variant Attributes
Allele Frequency: 1.00
Allele Count: 2
QD: 34.51
Mapping Quality: 59.41
SOR: 1.123
MLEAC: 2
ExcessHet: 0.0000
MLEAF: 1.00
Depth: 62
Total Alleles: 2
FS: 0.000

G.variants.varonl...

Chr: chr01
Position: 5414572
ID: .

Genotype Information
Sample: G
Genotype: T/T
Quality: 99
Type: HOM_VAR
Is Filtered Out: No

Genotype Attributes
AD: 0,59
Genotype Quality: 99
Depth: 59
PL: 2170,177,0

Figure above shows visualization of variant calling file using IGV V2.12.3. Detected variant is shown on the variant track. Details such as Quality score can be expanded by clicking on the nucleotide.Explanation of vcf fields can be found here: https://samtools.github.io/hts-specs/VCFv4.1.pdf

2.1.4 Building of distance tree based on vcf files

All filtered vcf files generated from each sample is first merged using bcftools (6). The vcf file is then converted to the PHYLIP format using vcf2phylip (7), and a neighbor joining tree was constructed using TreeBest (8) with 1000 bootstrapping.

References

1. Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing.

2. Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17:10–12.

3. Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.

4. Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100.

5. Yan MY, Ferguson B, Bimber BN. 2019. VariantQC: a visual quality control report for variant evaluation. Bioinformatics 35:5370–5371.

6. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. 2021. Twelve years of SAMtools and BCFtools. Gigascience 10:1–4.

7. Ortiz EM. 2019. vcf2phylip v2.0: convert a VCF matrix into several matrix formats for phylogenetic analysis.

8. GitHub - Ensembl/treebest: TreeBeST: Tree Building guided by Species Tree (Ensembl Compara modifications).