snakemake project for variant calling

python group

September 5, 2019

variant calling workflow using snakemake language

variant calling is the identification of nucleotide difference on a given reference genome or a transcriptome.Variant calling has become widely accepted in human genetics as a way of identifying variants associated with a specific trait, population or hereditary diesases. Using the standard pipeline available for identification of variant calling from H3ABionet community.We planned to make a more portable and reproducible workflow, for ease of analysis on any given platform. snakemake language offers this opportunity, due to its ability to work across different platform and it use of the python syntax, which is a pipelne language.

The pipeline was designed based on the standard operating procedures (SOPs) from H3Africa website The pipeline was divided into three phases : phase one: preprocesing of reads (fastqc analysis(fastqc), adapter removal and contaminate removale(trimmomatics)), Phasetwo:Intial variant discovery(alignment to reference genome (bwa and samtools), deduplication(sambamba), basequality score recalibaration (BSQR protocol from GATk)), Phase three: variant annotation and prioritization(SNP and INDEL variant prioritization(VSQR protocol from GATK)))

METHODOLOGY

To create the snakemake workflow:

Install either anaconda/bioconda platform

First set the environment for analysis: in our repo we already have given the instructions to follow: seting_up_the_environment

After setting the enviroment one can access the snakemake code from : from the repo

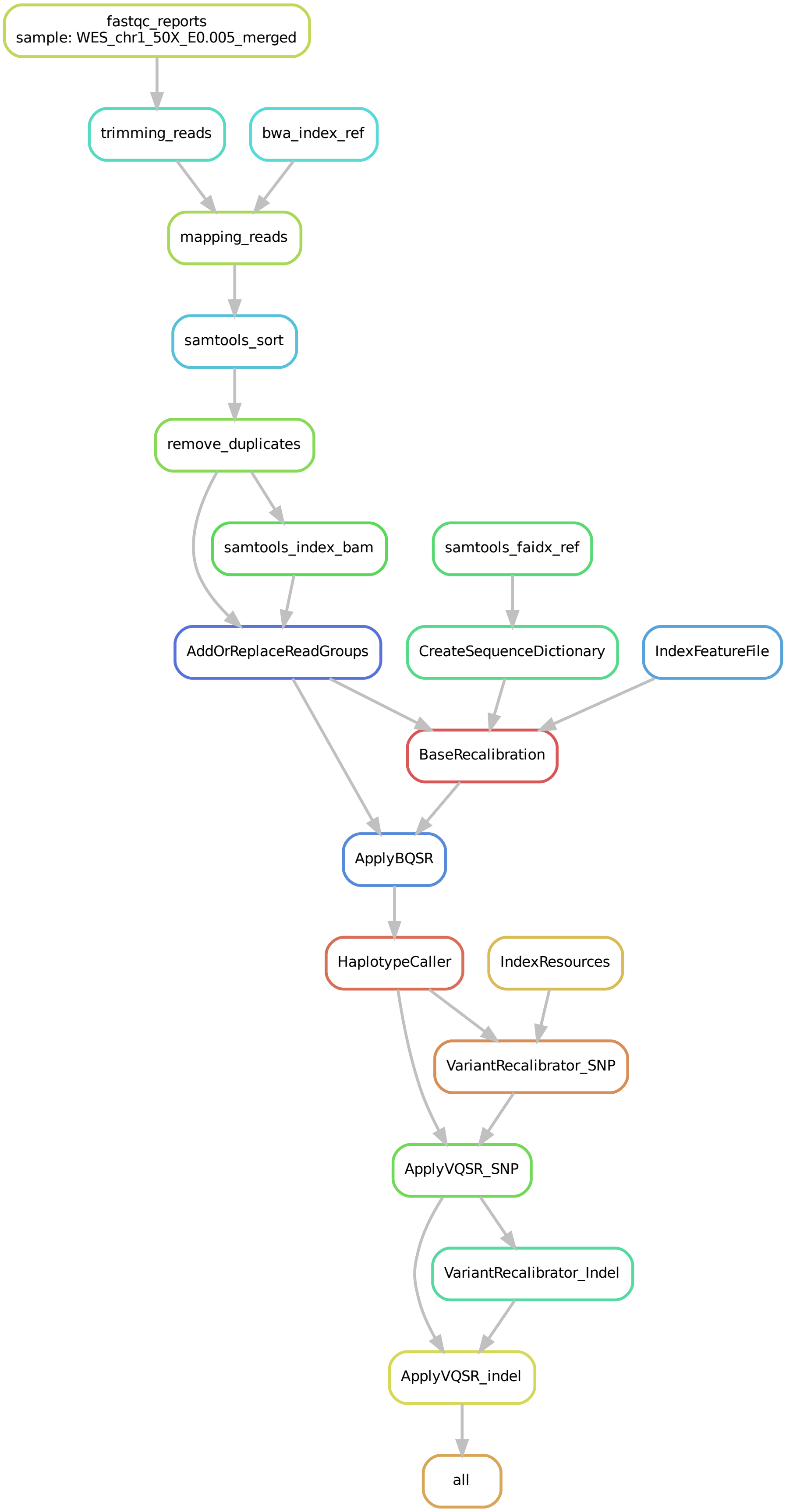
```r
# install packages(runnning conda command on r studio)
#install.packages("reticulate")
#install.packages("tidyverse")
# library installation
library(reticulate)
library(tidyverse)
```

```
## -- Attaching packages -------------------------------------------------------------------------------

## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.4.0

## -- Conflicts -----------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

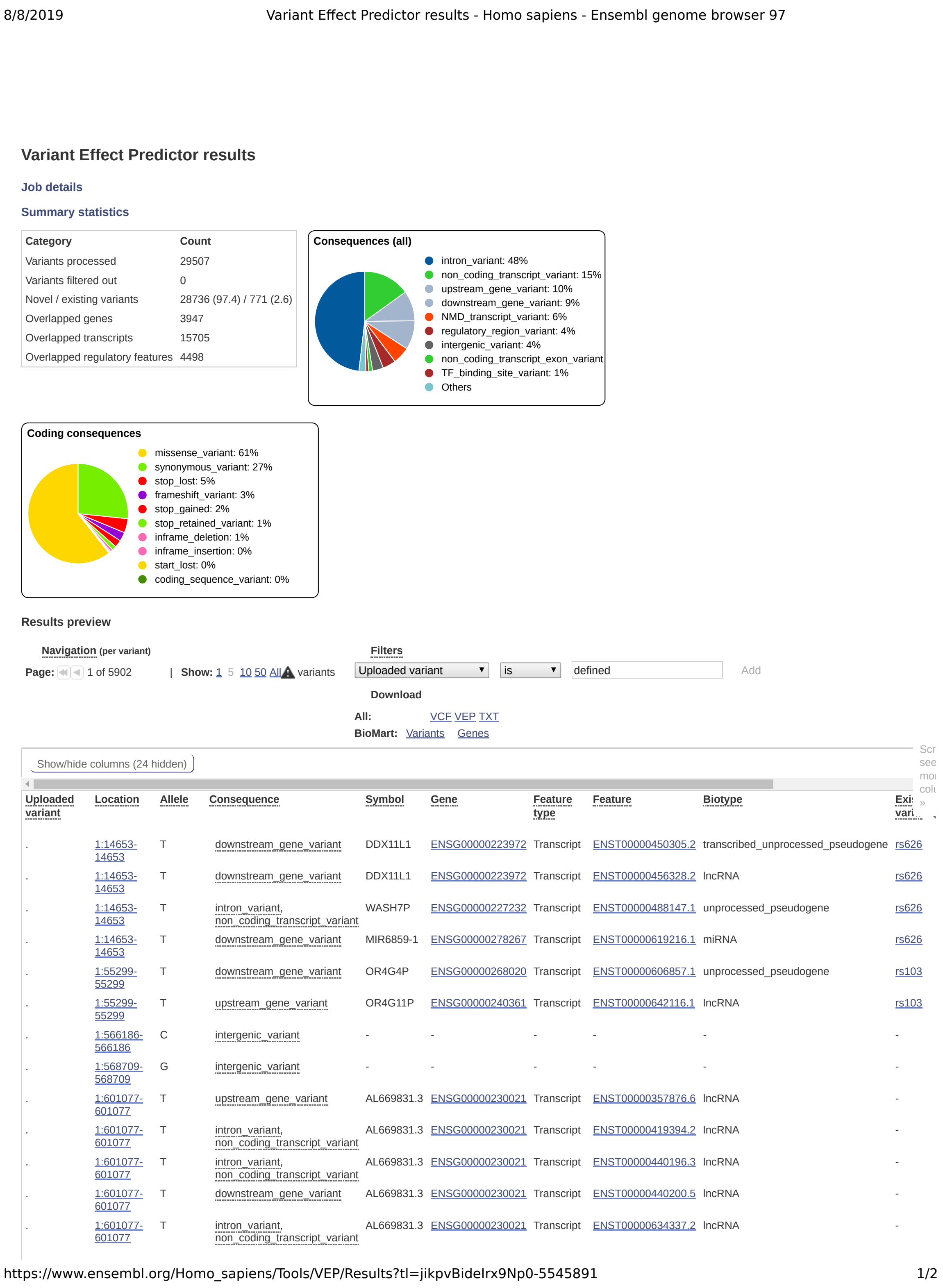
```r
# include python on the r script command
knitr::knit_engines$set(python = reticulate::eng_python)
```

```r
# set the environmnt to the directory with the snakmake file
## Set working directory.
setwd("/home/icipe/Variant_Calling_Project-/pipeline/")

#it is include the environments
conda_list()[[1]][1] %>%
  use_condaenv(required = TRUE)

# dry run of the snakemake command
# use of intern = true is used to display knitr output in either pdf or html
system("/home/icipe/miniconda3/envs/variant_calling/bin/snakemake -np")

#Running the command after confrimming with the dry run command

system("/home/icipe/miniconda3/envs/variant_calling/bin/snakemake")
```

```r
# library installation
library(reticulate)
library(tidyverse)
# displays the workflow directly
system("/home/icipe/miniconda3/envs/variant_calling/bin/snakemake --dag |dot |display " )
```


fastqc_reports
sample: WES_chr1_50X_E0.005_merged
trimming_reads
bwa_index_ref
mapping_reads
samtools_sort
remove_duplicates
samtools_index_bam
samtools_faidx_ref
AddOrReplaceReadGroups
CreateSequenceDictionary
IndexFeatureFile
BaseRecalibration
ApplyBQSR
HaplotypeCaller
IndexResources
VariantRecalibrator_SNP
ApplyVQSR_SNP
VariantRecalibrator_Indel
ApplyVQSR_indel
all

some of the variants identified from the work flow was analysed with the variant calling predictor effector from Embl database. Variant prector reports from our repo



8/8/2019 Variant Effect Predictor results - Homo sapiens - Ensembl genome browser 97

Variant Effect Predictor results

Job details

Summary statistics

Category	Count
Variants processed	29507
Variants filtered out	0
Novel / existing variants	28736 (97.4) / 771 (2.6)
Overlapped genes	3947
Overlapped transcripts	15705
Overlapped regulatory features	4498

Consequences (all)

- ● intron_variant: 48%
- ● non_coding_transcript_variant: 15%
- ● upstream_gene_variant: 10%
- ● downstream_gene_variant: 9%
- ● NMD_transcript_variant: 6%
- ● regulatory_region_variant: 4%
- ● intergenic_variant: 4%
- ● non_coding_transcript_exon_variant
- ● TF_binding_site_variant: 1%
- ● Others

Coding consequences

- ● missense_variant: 61%
- ● synonymous_variant: 27%
- ● stop_lost: 5%
- ● frameshift_variant: 3%
- ● stop_gained: 2%
- ● stop_retained_variant: 1%
- ● inframe_deletion: 1%
- ● inframe_insertion: 0%
- ● start_lost: 0%
- ● coding_sequence_variant: 0%

Results preview

Navigation (per variant) Filters

Page: 1 of 5902 | Show: 1 5 10 50 All variants Uploaded variant ▾ is ▾ defined Add

Download

All: VCF VEP TXT

BioMart: Variants Genes

Show/hide columns (24 hidden)

Scr see mor colu »

Uploaded variant	Location	Allele	Consequence	Symbol	Gene	Feature type	Feature	Biotype	Exi vari
.	1:14653-14653	T	downstream_gene_variant	DDX11L1	ENSG00000223972	Transcript	ENST00000450305.2	transcribed_unprocessed_pseudogene	rs626
.	1:14653-14653	T	downstream_gene_variant	DDX11L1	ENSG00000223972	Transcript	ENST00000456328.2	lncRNA	rs626
.	1:14653-14653	T	intron_variant, non_coding_transcript_variant	WASH7P	ENSG00000227232	Transcript	ENST00000488147.1	unprocessed_pseudogene	rs626
.	1:14653-14653	T	downstream_gene_variant	MIR6859-1	ENSG00000278267	Transcript	ENST00000619216.1	miRNA	rs626
.	1:55299-55299	T	downstream_gene_variant	OR4G4P	ENSG00000268020	Transcript	ENST00000606857.1	unprocessed_pseudogene	rs103
.	1:55299-55299	T	upstream_gene_variant	OR4G11P	ENSG00000240361	Transcript	ENST00000642116.1	lncRNA	rs103
.	1:566186-566186	C	intergenic_variant	-	-	-	-	-	-
.	1:568709-568709	G	intergenic_variant	-	-	-	-	-	-
.	1:601077-601077	T	upstream_gene_variant	AL669831.3	ENSG00000230021	Transcript	ENST00000357876.6	lncRNA	-
.	1:601077-601077	T	intron_variant, non_coding_transcript_variant	AL669831.3	ENSG00000230021	Transcript	ENST00000419394.2	lncRNA	-
.	1:601077-601077	T	intron_variant, non_coding_transcript_variant	AL669831.3	ENSG00000230021	Transcript	ENST00000440196.3	lncRNA	-
.	1:601077-601077	T	downstream_gene_variant	AL669831.3	ENSG00000230021	Transcript	ENST00000440200.5	lncRNA	-
.	1:601077-601077	T	intron_variant, non_coding_transcript_variant	AL669831.3	ENSG00000230021	Transcript	ENST00000634337.2	lncRNA	-

https://www.ensembl.org/Homo_sapiens/Tools/VEP/Results?tl=jikpvBideIrx9Np0-5545891 1/2

Snakemake is a diverse language, that can be used for manipulation of data, in this example: we would like to diplay only phase one of the variant calling analysis.

Employing rmarkdown and commands from snakemake we demonstrate the versatity of the workflow language

```
#Also one has the option to run any number of rules they require
# library installation
library(reticulate)
library(tidyverse)
# command from snakemake (diplays the output of phase one script: preprocesing of reads
#(fastqc analysis(fastqc), adapter removal and contaminate removal(trimmomatics)))

# to run a dry run of the snakemake command

# removing intern = true displays the results in the console
system("/home/icipe/miniconda3/envs/variant_calling/bin/snakemake -n --until trimming_reads")

# use of intern = true is used to display knitr output in either pdf or html
system("/home/icipe/miniconda3/envs/variant_calling/bin/snakemake -n --until trimming_reads", intern = T
```

```
##  [1] ""
##  [2] "rule fastqc_reports:"
##  [3] "    input: Data/reads/WES_chr1_50X_E0.005_merged_read1.fq.gz, Data/reads/WES_chr1_50X_E0.005_me
##  [4] "    output: analyses/fastqc/WES_chr1_50X_E0.005_merged_read1_fastqc.html"
##  [5] "    log: logs/fastqc/WES_chr1_50X_E0.005_merged.log"
##  [6] "    jobid: 1"
##  [7] "    benchmark: benchmarks/fastqc/WES_chr1_50X_E0.005_merged.txt"
##  [8] "    wildcards: sample=WES_chr1_50X_E0.005_merged"
##  [9] ""
## [10] ""
## [11] "rule trimming_reads:"
## [12] "    input: Data/reads/WES_chr1_50X_E0.005_merged_read2.fq.gz, analyses/fastqc/WES_chr1_50X_E0.0
## [13] "    output: analyses/trimmed/WES_chr1_50X_E0.005_merged_read2.paired.fastq.gz, analyses/trimmed
## [14] "    log: logs/trimming/WES_chr1_50X_E0.005_merged.log"
## [15] "    jobid: 0"
## [16] "    benchmark: benchmarks/trimming/WES_chr1_50X_E0.005_merged.txt"
## [17] "    wildcards: sample=WES_chr1_50X_E0.005_merged"
## [18] ""
## [19] "Job counts:"
## [20] "\tcount\tjobs"
## [21] "\t1\tfastqc_reports"
## [22] "\t1\ttrimming_reads"
## [23] "\t2"
```

```
#Also one has the option to run any number of rules they require

###library installation
library(reticulate)
library(tidyverse)

# this useful for the application of snakemake command
system("/home/icipe/miniconda3/envs/variant_calling/bin/snakemake -n --help")
```

The snakemake pipeline was evaluated with data used for acreditation in H3Africa consortium. The data was evaluated to have reported 27921 SNPs and 1589 INDELs demonstrating the functionality of our work flow. Although the snakemake workflow was able to generate the SNPs vcf files. Additional study is required for other types of variant studies.