SeqVItA
Sequence Variant Identification and Annotation Platform for Next Generation Sequencing Data

Tutorial

13 July 2018

Prashanthi Dharanipragada, Sampreeth Reddy Seelam and Nita Parekh
Center for Computational Natural Science and Bioinformatics
International Institute of Information Technology, Hyderabad, India

1. About SeqVItA

Sequence variants Identification and Annotation (SeqVItA) platform enables the user to identify sequence variants that include SNVs and short INDELs, in the whole genome, whole exome or targeted sequence data. SeqVItA identifies both germline and somatic sequence variants in NGS data. It is implemented in a combination of programming languages (C++, R, and Bash) and the variant calling step is based on Fisher exact test. Variant annotation and prioritization feature in SeqVItA is particularly useful in analyzing a patient's genomic profile and assist in making an informed decision about the treatment plan best suited to the patient, thus leading to precision medicine. SeqVItA has a modular framework (Figure 1) and the user may use the annotation module with VCF as input generated from any other variant caller.

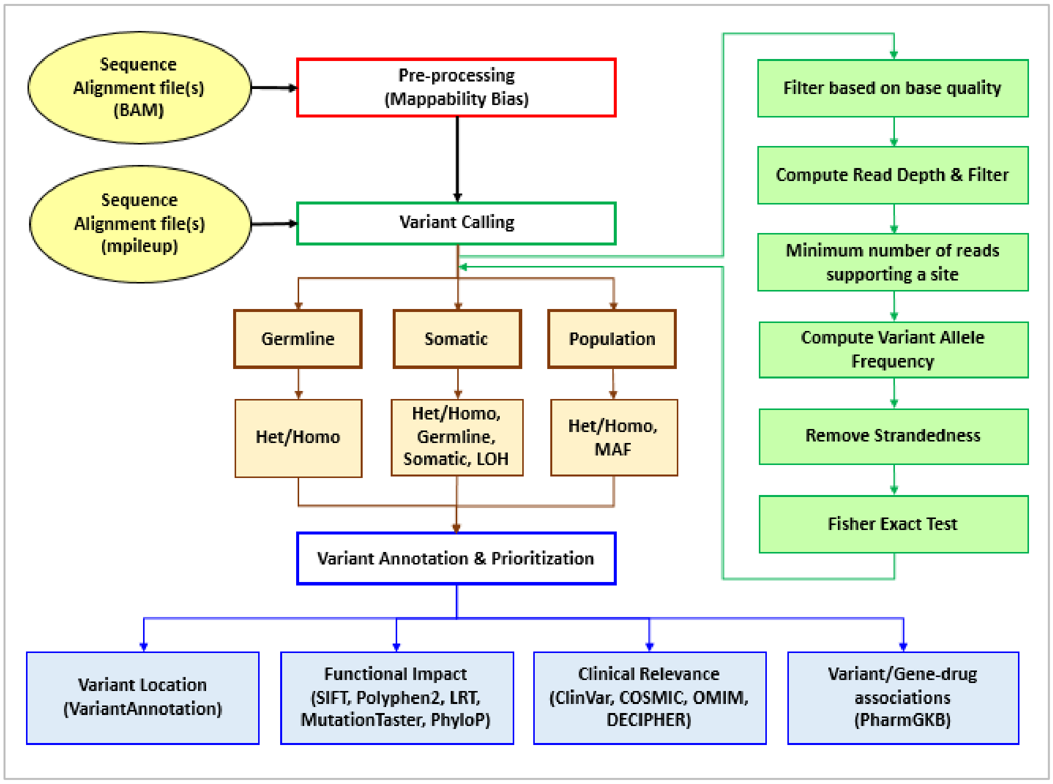

Sequence Alignment file(s) (BAM)
Sequence Alignment file(s) (mpileup)
Pre-processing (Mappability Bias)
Variant Calling
Germline
Somatic
Population
Het/Homo
Het/Homo, Germline, Somatic, LOH
Het/Homo, MAF
Variant Annotation & Prioritization
Filter based on base quality
Compute Read Depth & Filter
Minimum number of reads supporting a site
Compute Variant Allele Frequency
Remove Strandedness
Fisher Exact Test
Variant Location (VariantAnnotation)
Functional Impact (SIFT, Polyphen2, LRT, MutationTaster, PhyloP)
Clinical Relevance (ClinVar, COSMIC, OMIM, DECIPHER)
Variant/Gene-drug associations (PharmGKB)

Figure 1: Workflow of SeqVItA platform for identification and annotations of sequence variants from whole genome, whole exome or targeted sequence data

SeqVItA platform has a modular-framework with 3 main steps for SNV detection and annotation:

a) **Pre-processing**: This step in SeqVItA can be carried out only when the input file is an alignment file (in BAM format), for input file in mpileup format this step is skipped. Mapping quality recalibration (--Mqcorr) and filtering (--Mqread) are carried out in the step. The parameter values for these are user-defined. In our analysis, we observed that the recalibration of mappability scores is useful in reliable prediction of large INDELs (> 5 bp), however, it had no significant effect on the prediction accuracy of SNVs.

b) **Variant Calling *(variantCalling):*** From a given alignment file either in BAM format or mpileup format, SeqVItA predicts sequence variants and outputs in VCF file. Variant calling can be carried out in a single (***germline***), case-control (***somatic***) and multiple (***population***) samples using SeqVItA.

c) **Annotation *(annotate):*** For a given list of sequence variants in VCF format, annotate module helps in identifying biological significance of the variants through three categories of resources. These include (i) functional impact (SIFT, Polyphen2, MutationTaster, PhyloP and LRT scores), (ii) clinical/disease-associations (ClinVar, OMIM, COSMIC and DECIPHER) and (iii) Variant-drug associations (PharmGKB) and prioritizes the variants based on these three categories.

2. Installation

The pipeline is Linux-based and works with any latest version. SeqVItA can be used by installing from the source code, available at GitHub (https://github.com/Sampreeth13/seqvita).

SeqVItA from the source code

a. Download the following dependencies to run **SeqVItA** from the source code.

- R Dependencies:
  - Bioconductor packages:

  In R console, type the following:

```
>source("http://bioconductor.org/biocLite.R")
>biocLite("VariantAnnotation","rfPred", "SNPlocs.Hsapiens.dbSNP144.GRCh37")
```

  - R package: dplyr, tidyr

    In R console, type the following:

```
> install.packages("dplyr")
> install.packages("tidyr")
```

- Samtools (http://samtools.sourceforge.net/) (alternatively, in the terminal, type the following: sudo apt-get install samtools)
- OpenMP (https://www.openmp.org) (API for parallel computing, usually installed along with C/C++ in linux)

b. Download the source code by clicking ‘Download’ button from and then extract the zip file https://github.com/Sampreeth13/seqvita as shown below

```
unzip SeqVItA-master.zip
cd  SeqVItA-master
```

c. Download the annotations folder from the SeqVItA website (http://bioinf.iiit.ac.in/seqvita/) and add the contents of the folder into SeqVItA-master folder. Downloading annotations folder may take time due to large sized annotation files for hg19 reference genome assembly.

3. Workflow

We demonstrate the usage of our pipeline by considering a targeted exome sequence of cancer samples (Chromosome 1) mapped to hg19 human genome reference assembly. The test files can be downloaded from the SeqVItA website (http://bioinf.iiit.ac.in/seqvita/).

Table 1: Parameters considered for the detection of SNVs and INDELs in SeqVItA

Parameter	Description	Value
--Mqread	Mapping quality Cut-off (Only when alignment in BAM format is used as input)	20
--Mqcorr	Mapping quality correction using Samtools (Only when alignment in BAM format is used as input)	0
--Qbase	Minimum base quality at a position to count a read	15
--RD_th	Minimum read depth at a position to make a call	10
--VAR_th	Minimum supporting reads at a position to call alternate allele (variant)	2
--VAF_th	Minimum variant allele frequency threshold	0.20
--Strand_Bias	Minimum frequency to call homozygote	0.75
--p-value	Default p-value threshold for calling variants	0.01
--VAF_homo	Ignore variants with > 90% support from one strand	1
--somatic-p-value	p-value cut-off for calling somatic and LOH variants (Only used in somatic module)	0.05

(i) Germline Variant Calling from an Alignment File

Files required: alignment file (BAM), fasta file

Usage:

```
variantCalling -v SNP -ib <input bam> -r <ref genome> [Options] -o <output prefix>
variantCalling -v INDEL -ib <input bam> -r <ref genome> [Options] -o <output prefix >
variantCalling -v germline -ib <input bam> -r <ref genome> [Options] -o <output prefix >
```

Output: A VCF file with sequence variants

Sequence variants constitute SNVs and short INDELs. SeqVItA offers the user to either predict SNVs or INDELs separately using 'SNP' or 'INDEL' functions, respectively or can be simultaneously called using 'germline' function where both the types of sequence variants are reported in a single file. The input of these commands is an aligned file. The performance of SeqVItA is efficient in using a high-quality alignment file (low quality reads filtered and adopter trimmed) and preferably PCR duplicates marked. On using alignment files in BAM formats, users will have a choice to pre-process the data based on mapping quality of the reads. Parameters '--Mqbase' and '--Mqcorr' are considered in this step. Using Samtools, if '--Mqbase' is assigned with an integer value, mapping quality correction takes place using the following expression.

$$Mq' = \sqrt{\frac{Int - Mq}{Int}} \times Int$$

Here Int is a user defined integer '--Mqcorr' (default: 0) and Mq is the phred-scaled probability of a read being misaligned. The recalibrated file is generated in an mpileup format for further analysis. If the alignment has been obtained using only uniquely mapped reads, one may skip this step. Ideally, for alignment files generated using BWA and Bowtie2, 50 is an appropriate value to adjust the mapping qualities of the mismatched reads. This is followed by filtering of low mapping quality reads and the value of '--Mqbase' (default: 20) is user-defined depending on the aligner used. The pre-processing step in SeqVItA is to be used only when the input file is alignment file (in BAM format), for input file in mpileup format this step is skipped. A reference file has to be supplied and reference genome sequence for hg19 is made available in annotations folder. User may change any of the parameters listed in Table 1, else variant calling is carried out with default parameters. A VCF file is generated at the end of using these commands.

```
variantCalling -v germline -ib Test1.bam -r annotations/hg19.fa --Mqcorr 50 -o Test1_output
```

The above command results in Test1_output.vcf, a VCF file with both SNVs and short INDELs identified in Test alignment file.

(ii) Germline variant calling from mpileup file

Files required: Mpileup, reference genome sequence file (Fasta)

Usage:

```
variantCalling -v SNP -im <input mpileup> -r <ref genome> [Options] -o <output prefix>
variantCalling -v INDEL -im <input mpileup> -r <ref genome> [Options] -o <output prefix>
variantCalling -v germline -im <input mpileup> -r <ref genome> [Options] -o <output prefix>
```

Output: A VCF file with sequence variants

Similar to variant calling using BAM files, SeqVItA can also predict sequence variants using mpileup file. Mpileup file can be generated using Samtools. SeqVItA expects the input mpileup files to be of a high quality (pre-processing and post-alignment processing such as filtering low mapping quality sequence).

```
variantCalling -v germline -im Test1.mpileup -r annotations/hg19.fa -o Test_mpileup.vcf
```

The above command uses the mpileup file for prediction of both SNVs and INDELs. Please note the results obtained using this function may vary compared to directly using alignments file (BAM) due to differences in post-alignment processing carried out.

(iii) Somatic Variant Calling in Case-control Tumor Samples

Files required: alignment files (BAM or mpileup) for normal and tumor, reference genome sequence file (Fasta)

Usage:

```
variantCalling -v somatic --normal <normal bam> --tumor <tumor bam> -r <ref genome> [options] –o <output prefix>
variantCalling -v somatic -im <normal-tumor mpileup> [options] –o <output prefix>
```

Options:
--normal-read-depth: Minimum coverage threshold cutoff for the normal sample [Default: 8]
--tumor-read-depth: Minimum coverage threshold cutoff for tumor sample [Default: 6]

Output: A VCF file with somatic, germline, LOH and unknown variants

Somatic variant calling is carried out in SeqVItA with a matched-control sample of the tumor. The user may either provide a pair of aligned (bam) files or may directly give a combined mpileup file generated using Samtools with normal sequences followed by tumor sequences. There are two parameters for somatic function *viz.*, '--normal-read-depth' and '--tumor-read-depth' which are the minimum number of reads that have to be considered for variant calling in normal and tumor samples, respectively. The parameter '--RD_th' is not valid and all other values mentioned in Table 1 can be used. If the *p*-value for a base is smaller than the threshold cut-off (--somatic-p-value: 0.05) and the normal matches the reference allele, the base is classified as 'Somatic' and it is classified as 'LOH' if the normal is heterozygous and homozygous in the tumor sample. In case the *p*-value is greater than the threshold, the *p*-value is recomputed by combining tumor and normal read counts for each allele and the base is considered as 'germline'.

```
variantCalling -v somatic --normal Test2_normal.bam --tumor Test2.bam -r
annotations/hg19.fa -o Test2_somatic
```

The above command results in a VCF file with Somatic, Germline and LOH variants. For each of them, SeqVItA computes somatic p-value and variant p-value. SeqVItA also offers s***plitVCF*** function (with ***'--somatic'*** option) that enables the user to separate VCF file generated from the somatic module into four files based on the type of the sequence variant predicted *i.e.*, into somatic, germline, LOH and unknown.

```
splitVCF --somatic -i Test2_somatic.vcf -o Test2_splitVCF
```

The above command splits the Test2_somatic.vcf into four files *viz.*, Test2_splitVCF_somatic.vcf, Test2_splitVCF_LOH.vcf, Test2_splitVCF_germline.vcf and Test2_splitVCF_unknown.vcf. Additionally, SeqVItA also gives 'High' and 'Low' classification of variants based on variant allele frequency and p-value computed above. A somatic variant is reported as 'High' priority if VAF ≥ 10% in tumor, < 5% in normal and *p*-value < 0.07, and LOH is reported as 'High' priority if VAF ≥ 10% in normal and *p*-value < 0.07. Germline variants are described as 'High' priority if VAF ≥ 10% in both tumor and normal samples. Any sequence variant not meeting these criteria is assigned 'Low' priority. The function ***splitVCF*** (with ***'--confidence'*** option) can also be used to filter high confidence sequence variants predicted in this module.

```
splitVCF --confidence -i Test2_somatic.vcf -o Test2_high
```

The above command generates a VCF file, Test2_high.vcf, with only high confidence sequence variants predicted in Test2_somatic.vcf file.

(iv) Variant Calling across Multiple Population samples

Files required: Multiple alignment files (BAM) or single mpileup file generated from multiple files, reference genome (Fasta format)

Usage:

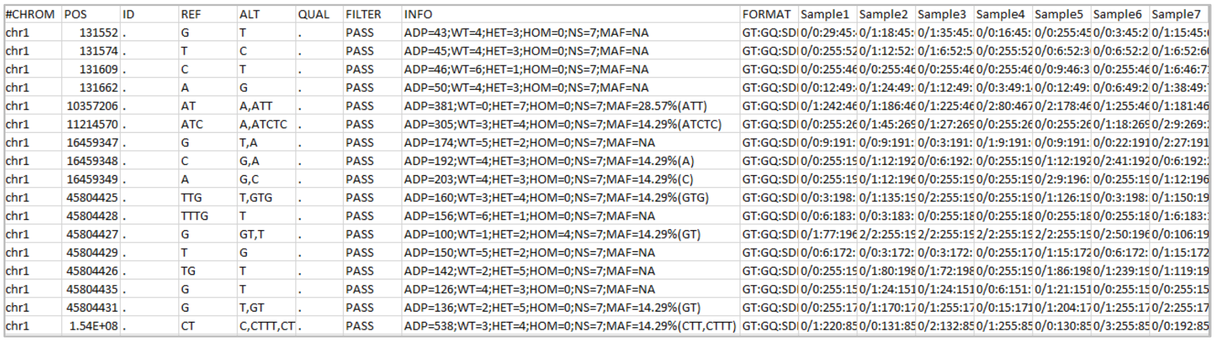
variantCalling -v population -ib <input bam 1> <input bam 2> <input bam 3> <input bam 4> … -r <ref genome> [Options] -o <output prefix >
variantCalling -v population -im <single mpileup generated from multiple samples> [Options] -o <output prefix >

Output: A VCF file with sequence variants identified across multiple samples

SeqVItA offers identification of rare and common alleles across multiple population samples. The input can be several alignment files in BAM format or a single mpileup file generated from multiple BAM files using Samtools mpileup module. The procedure followed in ***population*** module is similar to the ***germline*** module. The sequence variant detection and *p*-value computation are carried out individually for each sample using the one-tailed Fisher Exact Test. Additionally, in this module, SeqVItA reports the Minor Allele Frequency (MAF) for each variant. It is defined as the frequency of the second most common allele identified in the given population dataset. Based on MAF value, user may categorize the predicted sequence variant as 'rare' (MAF < 0.01) or 'common' (MAF > 0.01). Following command is used to identify sequence variants across 5 alignment files.

variantCalling -v population -ib Test1.bam Test2.bam Test3.bam Test4.bam Test5.bam -r annotations/hg19.fa -o Test_population

This results in Test_population.vcf, where information about each sample is sequentially summarized from the 10th column. A sequence variant may have more than one allele identified in multiple samples and these alleles are reported under 'ALT' column, as shown below Figure 2. Samples with no sequence variant are represented by 0/0 and 0/1 if the sample contains one variant allele. This is reported in the respective sample information columns (Figure 2). In case more than one alternate allele is identified in any sample, then it is represented as 0/2 under that sample and both the alleles are reported in the 'ALT' column (comma separated) in VCF file. MAF is reported with percentage frequency of the allele (given in the parenthesis) in the 'INFO' column, MAF=NA implies no minor allele is identified.



#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7
chr1	131552	.	G	T	.	PASS	ADP=43;WT=4;HET=3;HOM=0;NS=7;MAF=NA	GT:GQ:SD	0/0:29:45:	0/1:18:45:	0/1:35:45:	0/0:16:45:	0/0:255:45	0/0:3:45:2	0/1:15:45:
chr1	131574	.	T	C	.	PASS	ADP=45;WT=4;HET=3;HOM=0;NS=7;MAF=NA	GT:GQ:SD	0/0:255:52	0/1:12:52:	0/1:6:52:5	0/0:255:52	0/0:6:52:3	0/0:6:52:2	0/1:6:52:6
chr1	131609	.	C	T	.	PASS	ADP=46;WT=6;HET=1;HOM=0;NS=7;MAF=NA	GT:GQ:SD	0/0:255:46	0/0:255:46	0/0:255:46	0/0:255:46	0/0:9:46:3	0/0:255:46	0/1:6:46:7
chr1	131662	.	A	G	.	PASS	ADP=50;WT=4;HET=3;HOM=0;NS=7;MAF=NA	GT:GQ:SD	0/0:12:49:	0/1:24:49:	0/1:12:49:	0/0:3:49:1	0/0:12:49:	0/0:6:49:2	0/1:38:49:
chr1	10357206	.	AT	A,ATT	.	PASS	ADP=381;WT=0;HET=7;HOM=0;NS=7;MAF=28.57%(ATT)	GT:GQ:SD	0/1:242:46	0/1:186:46	0/1:225:46	0/2:80:467	0/2:178:46	0/1:255:46	0/1:181:46
chr1	11214570	.	ATC	A,ATCTC	.	PASS	ADP=305;WT=3;HET=4;HOM=0;NS=7;MAF=14.29%(ATCTC)	GT:GQ:SD	0/0:255:26	0/1:45:269	0/1:27:269	0/0:255:26	0/0:255:26	0/1:18:269	0/2:9:269:
chr1	16459347	.	G	T,A	.	PASS	ADP=174;WT=5;HET=2;HOM=0;NS=7;MAF=NA	GT:GQ:SD	0/0:9:191:	0/0:9:191:	0/0:3:191:	0/1:9:191:	0/0:9:191:	0/0:22:191	0/2:27:191
chr1	16459348	.	C	G,A	.	PASS	ADP=192;WT=4;HET=3;HOM=0;NS=7;MAF=14.29%(A)	GT:GQ:SD	0/0:255:19	0/1:12:192	0/0:6:192:	0/0:255:19	0/1:12:192	0/2:41:192	0/0:6:192:
chr1	16459349	.	A	G,C	.	PASS	ADP=203;WT=4;HET=3;HOM=0;NS=7;MAF=14.29%(C)	GT:GQ:SD	0/0:255:19	0/1:12:196	0/0:255:19	0/0:255:19	0/2:9:196:	0/0:255:19	0/1:12:196
chr1	45804425	.	TTG	T,GTG	.	PASS	ADP=160;WT=3;HET=4;HOM=0;NS=7;MAF=14.29%(GTG)	GT:GQ:SD	0/0:3:198:	0/1:135:19	0/2:255:19	0/0:255:19	0/1:126:19	0/0:3:198:	0/1:150:19
chr1	45804428	.	TTTG	T	.	PASS	ADP=156;WT=6;HET=1;HOM=0;NS=7;MAF=NA	GT:GQ:SD	0/0:6:183:	0/0:3:183:	0/0:255:18	0/0:255:18	0/0:255:18	0/0:255:18	0/1:6:183:
chr1	45804427	.	G	GT,T	.	PASS	ADP=100;WT=1;HET=2;HOM=4;NS=7;MAF=14.29%(GT)	GT:GQ:SD	0/1:77:196	2/2:255:19	2/2:255:19	2/2:255:19	2/2:255:19	0/2:50:196	0/0:106:19
chr1	45804429	.	T	G	.	PASS	ADP=150;WT=5;HET=2;HOM=0;NS=7;MAF=NA	GT:GQ:SD	0/0:6:172:	0/0:3:172:	0/0:3:172:	0/0:255:17	0/1:15:172	0/0:6:172:	0/1:15:172
chr1	45804426	.	TG	T	.	PASS	ADP=142;WT=2;HET=5;HOM=0;NS=7;MAF=NA	GT:GQ:SD	0/0:255:19	0/1:80:198	0/1:72:198	0/0:255:19	0/1:86:198	0/1:239:19	0/1:119:19
chr1	45804435	.	G	T	.	PASS	ADP=126;WT=4;HET=3;HOM=0;NS=7;MAF=NA	GT:GQ:SD	0/0:255:15	0/1:24:151	0/1:24:151	0/0:6:151:	0/1:21:151	0/0:255:15	0/0:255:15
chr1	45804431	.	G	T,GT	.	PASS	ADP=136;WT=2;HET=5;HOM=0;NS=7;MAF=14.29%(GT)	GT:GQ:SD	0/0:255:17	0/1:170:17	0/1:255:17	0/0:15:171	0/1:204:17	0/1:255:17	0/2:255:17
chr1	1.54E+08	.	CT	C,CTTT,CT	.	PASS	ADP=538;WT=3;HET=4;HOM=0;NS=7;MAF=14.29%(CTT,CTTT)	GT:GQ:SD	0/1:220:85	0/0:131:85	0/2:132:85	0/1:255:85	0/0:130:85	0/3:255:85	0/0:192:85

Figure 2: Snapshot of SeqVItA output file on using *population* module across multiple data samples

(v) Variant calling from Whole exome or Targeted sequencing data

Files required: alignment file (BAM), reference genome sequence file (Fasta), coordinate file (BED)

Usage:

```
variantCalling -v SNP -ib <input bam> -r <ref genome> --bed <Coordinate_file> [Options] -o <output prefix>
variantCalling -v INDEL -ib <input bam> -r <ref genome> --bed <Coordinate_file> [Options] -o <output prefix>
variantCalling -v germline -ib <input bam> -r <ref genome> --bed <Coordinate_file> [Options] -o <output prefix>
variantCalling -v somatic --normal <normal bam> --tumor <tumor bam> -r <ref genome> --bed <Coordinate_file> [Options] -o <output prefix>
variantCalling -v population -ib <input bam 1> <input bam 2> <input bam 3> <input bam 4> … -r <ref genome>  --bed <Coordinate_file> [Options] -o <output prefix >
```

Output: A VCF file with sequence variants in only given coordinates

SeqVItA identifies sequence variants in whole genome, whole exome and targeted sequence data. User can provide a list of locations (a tab-separated coordinate file) and SeqVItA will accurately predict the sequence variants within the given locations. This avoids false positives and increases the speed of the sequence variants prediction.

(vi) Variant annotation

Files required: variant file (VCF) generated from SeqVItA or any other variant caller

Usage:

```
annotate -i <input.vcf>  --geneBasedDrug -o <Output prefix>
```

Output: A tab-separated file with all the annotations available in SeqVItA

Along with detection of sequence variants in NGS data, SeqVItA also offers annotation of the predicted variants. The modular framework of SeqVItA enables the user to annotate not only the variants predicted by SeqVItA but also from various other variant callers such as GATK, VarScan2 etc. Information on location and type of sequence variants, genes spanning them and dbSNP id if available are reported. In addition to these, three categories of annotations are available *viz*., (i) functional impact (SIFT, Polyphen2, MutationTaster, PhyloP and LRT scores), (ii) clinical/disease-associations (ClinVar, OMIM, COSMIC and DECIPHER) and (iii) Variant-drug associations (PharmGKB). Based on these three categories, the sequence variants are also prioritized and each of the variants is categorized into ‘High’, ‘Medium’ and ‘Low’ priority. A variant is assigned ‘High’ priority if the variant has a high score (> 0.65) in any one of the five resources (SIFT, Polyphen2, MutationTaster, LRT and PhyloP) based on functional impact, has clinical association identified in at least one of the three resources (ClinVar, COSMIC and OMIM), and has variant-drug association annotated in PharmGKB. A variant is assigned ‘Medium’ priority if it has reported significance from any of these resources. Any variant not meeting the above criteria is assigned ‘low’ priority.

```
annotate -i Test.vcf -o Test_annotated
```

Using the above command, SeqVItA generates an annotated file, Test_annotated, for variants reported in Test.vcf. SeqVItA also offers gene-drug associations which is a less stringent condition, compared to the default one which maps the dbSNP to a drug in PharmGKB. For gene-drug association, the user needs to add ‘-d’ and ‘--genebased’ parameter while using annotate module.

```
annotate -i Test.vcf -d --genebased -o Test_geneBasedDrug_annotated
```

The above results in an annotated file, Test_geneBasedDrug_annotated, with genes-drug associations identified in PharmGKB.