VaRank Manual

Version 1.0
VaRank is a program for genetic Variant Ranking from NGS data

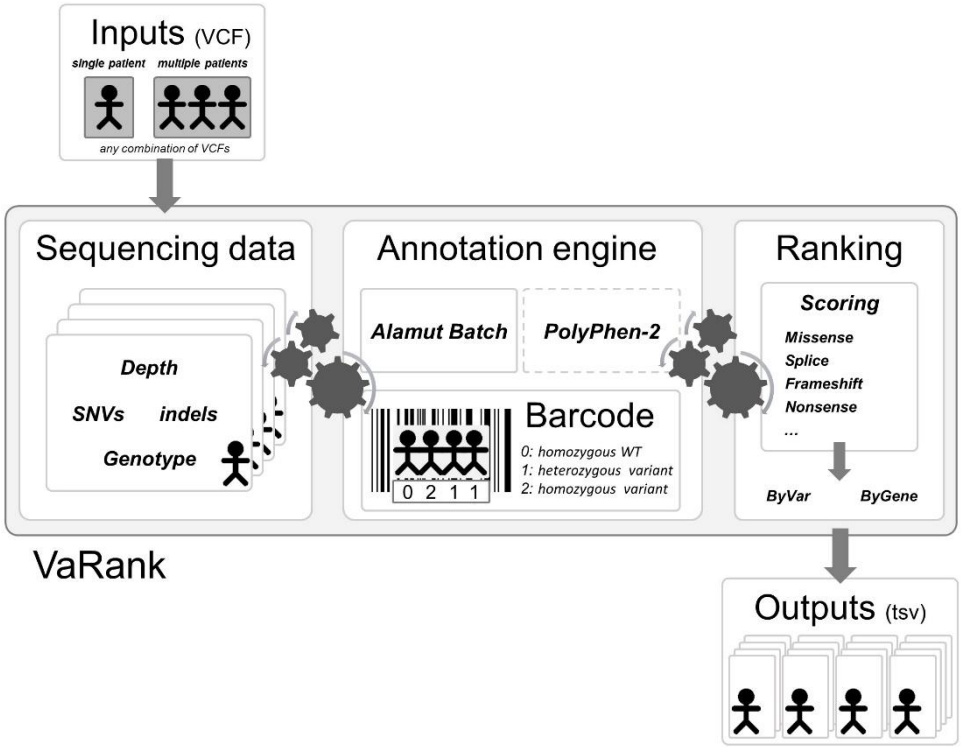
Please feel free to contact us for any suggestions or bug reports
email: veronique.geoffroy@inserm.fr; jeanmuller@unistra.fr

==========
TABLE OF CONTENTS
==========

==========

1. INTRODUCTION
==========

VaRank is a program designed for variant ranking from next generation sequencing data. It provides a comprehensive workflow for annotating and ranking SNVs and indels.
Four modules create the strength of this workflow:
(i) Quality scores summary (total and variant depth of coverage, phred like information), to filter out false positive calls.
(ii) Alamut Batch annotations, to integrate genetic and predictive information (functional impact, putative effects in the protein coding regions, population frequency...) from different sources, using HGVS nomenclature.
(iii) Barcode representing the presence/absence of variants (with homozygote/heterozygote status), to search for recurrence between families or group of individuals.
(iv) Prioritization score, to rank variants according to their predicted pathogenic status.



2. INSTALLATION/REQUIREMENTS
==========

The VaRank program is written in the Tcl/Tk language. Modern Unix systems have this already installed (or can be downloaded from http://www.tcl.tk/).This algorithm is composed of different other programs and databases:

- VaRank sources can be downloaded from http://www.lbgi.fr/VaRank under the GNU GPL license.
- Alamut Batch (Interactive Biosoftware, Rouen, France), if you do not own a license, a 30-day free trial can be requested here (http://www.interactive-biosoftware.com/request-trial-alamut/)

Optional:
- PolyPhen-2 (PPH2) provides prediction of functional effects of human nsSNPs (Adzhubei IA et al. Nat Methods 2010). It needs to be locally installed to be used. You can freely download it from http://genetics.bwh.harvard.edu/pph2/dokuwiki/downloads

- Protein databases can be used to connect to PPH2. UniProt and RefSeq can respectively be downloaded from:
ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/proteomes/
ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/mRNA_Prot/human.protein.faa.gz

The protein databases files help VaRank to extract the protein sequences and 1/ check the aminoacid change to be tested and 2/ submit the protein sequence to PPH2 if no accession are precomputed.

The source .tar.gz should be extracted and uncompressed to any directory. The installation requires simply to set the following environment variables:
- $VARANK : VaRank installation directory
- $ALAMUT : Alamut Batch installation directory

The following environment variable is optional:
- $PPH : PolyPhen-2 installation directory

By default the VaRank installation directory looks like this:

```
VaRank                  #The program installation directory
 |
 |----- bin/            #Where an alias is set to the main .tcl script
 |
 |----- DataBases/      #Where to store the UniProt and RefSeq fasta files
 |
 |----- sources/        #Where the .tcl files are stored
 |
 |----- configfile      #an example of configfile that can be copied to any analysis director
 |                      #for modification purpose
 |
 |----- help.txt        #description of VaRank options and default settings
 |
 |----- License.txt     #GNU GPL license
 |
 |----- README          #This file
```

Make sure the program find correctly the Tcl interpreter, by default the best way to make a Tcl script executable is to put the following as the first line of the main script (which is already done in VaRank-main.tcl):

```
#!/usr/bin/env tclsh
```

But it can be changed to any other path like:

```
#!/usr/local/ActiveTcl/bin tclsh
```

Typically, you can create an alias of the main Tcl script “sources/VaRank-main.tcl” for example to “VaRank”, place it in the “/bin” directory” (this is done be default already) and add the path to this in your $PATH.

3. INPUT
=======

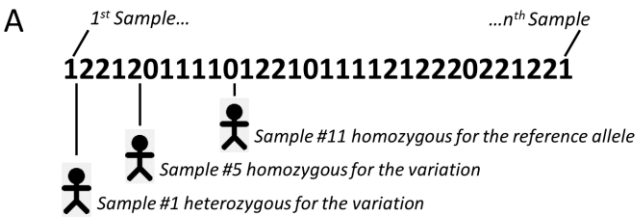
VaRank supports the commonly used VCF (Variant Call Format, https://github.com/samtools/hts-specs) input format for variants analysis that allows the program to be easily integrated into NGS bioinformatics analysis pipelines.

VaRank takes also several argument as options to the command line that are detailed in section 6 ("USAGE / OPTIONS"). The different arguments can be passed either on the command line or using a specific file named "configfile" that needs to be put in the same directory as the input VCF files. An example of configfile is provided in the VaRank installation directory.

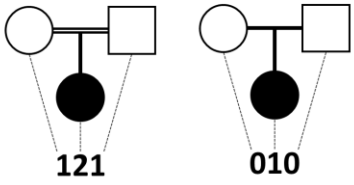
a. Family Barcode

The barcode in VaRank allows a quick overview of the presence/absence status of each variant and their zygosity status within the analyzed individuals ("0" representing homozygous wild type, "1" heterozygous and "2" homozygous for the variant, see the figure below Panel A). Panel B displays 3 variants example and 32 patients analyzed together.

A



C



B

Gene	Chr	Start	Ref	Mut	Zygosity	TotalRead Depth	VarRead Depth	cNomen	pNomen	Barcode	#Hom	#Het	#Allele	#Sample
BBS2	16	56548501	C	T	hom	142	142	c.209G>A	p.=	**22222222202222222222222222222222**	31	0	62	32
ALMS1	2	73716993	-	C	hom	143	126	c.7911dup	p.Asn2638Glnfs*24	**00000000000000000000000000200000**	1	0	2	32
TTC21B	2	166797646	C	T	het	144	144	c.601G>A	p.Val201Met	**12212011110122101111212220221221**	13	16	42	32

Together with the main barcode describing all the patients analyzed together in one VaRank run, one can define a second barcode. This second barcode named "familyBarcode" can be configured by the user to group selected samples (e.g. trios where affected child and parents could be specifically grouped together). This can be configured in the configfile by simply grouping sample names together. As an example, 2 families where the fam1 corresponds to a trio sequencing (proband and parents, see Figure Panel C) and fam2 with 2 affected child.

fam1: Sample1 Sample2 Sample3
fam2: Sample4 Sample5

Grouping sample names together allows also to follow the same naming convention for the files with the same prefix (fam1_ for all family members).

b. External Gene annotation

In order to further enrich the annotation for each variant and each gene, VaRank can integrate (using the option -extann) external annotations imported from a tab separated values file into the output files. The file format is easy and should look like this (1st line is a header including a column entitled "Gene" that should be the 1st column too). The following example has been set to provide annotation for the gene including the transmission mode of the gene (here AR means “autosomic recessive”), the number of missense and truncating mutations reported as well as the OMIM identifier.

Gene	Transmission	#Missense	#Truncating	Omim
ACY1	AR	4	2	104620
ADSL	AR	7	1	608222

4. OUTPUT

=========

VaRank provides 4 .tsv (TAB separated values) output files divided into 2 categories:

-Files named with "ByVar" contains variations sorted from the most to the least pathogenic (according to the VaRank score)

-Files named with "ByGene" contains variations classified by gene ("ByGene") where the list is sorted using the gene as a proxy to the score. Each gene is scored according to most pathogenic variant (homozygous) or the first two most pathogenic variants. In order to make sure that no variants are missed all gene variation are reported also below the variant(s) used to score the gene. This file is more suitable when dealing with a recessive mode of inheritance.

It is to notice that given the focus on genes in those output files, variants that could be attributed to several genes are duplicated and associated to each gene individually.

A part from these 2 categories, each file is also available in 2 versions:

-Raw file ("allVariants") with no variants filtered out.

-Already prefiltered files ("filteredVariants") with variants filtered out using the following criteria:

The default filters remove variants:

- -with a total depth of coverage <= 10x
- -with a supporting reads count <= 10x
- -with a percent of supporting reads <= 15%
- -with validated annotation in the dbSNP database (i.e. at least with 2 evidences) that are not pathogenic (from the ClinicalSignificance field in dbSNP)
- -with an allele frequency > 1% (extracted from the dbSNP database or the Exome Variant Server)

The “filtered” files can be considered as very stringent filtering step to ensure a very quick first analysis of the data. Users can always adapt the options to make fit his situation.

The output organization can be described as follows:

```
VcfDirectory
 |
 |----- configfile                          #if present can be used to define sample group and set
options of VaRank
 |
 |----- *InputFile.vcf                      #Input files
 |
 |----- Alamut/                             #Contains all Alamut Batch related files
 |      |----- AlamutInputFile_all.txt      #Alamut input file generated from the vcf(s) files
 |      |----- AlamutAnnotations_all.txt    #Alamut output file with annotated variants
 |      |----- AlamutUnnanotated_all.txt    #Alamut output file with unannotated variants
 |      |----- AlamutOutput_all.txt         #Alamut log file
 |
 |----- PPH2/                               #(option) Contains all PolyPhen-2 related files
 |      |----- PPH2input_all.txt            #PPH2 input file
 |      |----- PPH2features_all.txt         #PPH2 output file
 |      |----- PPH2humvar_all.txt           #PPH2 output file
 |      |----- PPH2errors_all.txt           #PPH2 log file
 |
 |----- fam#_SampleName_allVariants.rankingByVar.tsv
 |----- fam#_SampleName_filteredVariants.rankingByVar.tsv
 |
 |----- fam#_SampleName_allVariants.rankingByGene.tsv
 |----- fam#_SampleName_filteredVariants.rankingByGene.tsv
 |
 |----- fam#_SampleName_statistics.txt      #Short counts report (e.g. homozygous, heterozygous
 |                                          #and total counts) for each of the variant categories
 |
 |----- SNV_global_statistics.txt           #Contains the same counts as defined for each patient
 |                                          #but for the whole analyzed cohort
```

It is to notice that when no annotation is available for a specific column, the empty value is set to "NA". Exception is made for several numerical columns (including rsMAF, espEAMAF, espAAMAF, espAllMAF) where "-1" is used that allows the user to further filter information without losing data.

5. SCORING

VaRank uses the variation type (i.e. substitution, deletion, insertion, duplication) and the coding effect to score. The VaRank scoring is categorized from the most likely to the less likely pathogenic state as follows (score into parenthesis): known mutation (110), nonsense (100), frameshift (100), essential splice site (2 first bases before and after the exon) (90), start loss (80), stop loss (80), intron-exon boundary (donor site is -3 to +6, acceptor site -12 to +2) (70), missense (50), in-frame (40), deep intronic changes (25) and synonymous coding (10). Each category is further described in the USAGE/OPTIONS section and each score can be changed.

Each specific variant score is further adjusted using additional information. For this, variants are assessed at the genomic level (PhasCons) and at the protein level (SIFT and if installed PolyPhen-2), and an adjustment score (0 to +10) is added to the relevant category.

The scores in red reflect the range of score while the adjustment score is applied.
[1]Each variant score is adjusted (+5) if a conservation at the genomic level is observed (PhastCons cutoff >0.95)
[2]Missenses scores are adjusted (+10) for each deleterious prediction (SIFT and/or PPH2)

Variant Category	Options	VaRank Score	Definitions

Known mutation	S_Known	110	Known mutation as annotated by HGMD and/or dbSNP (rsClinicalSignificance="pathogenic/probable-pathogenic")
Nonsense	S_Nonsense[1]	100, 105	A single-base substitution in DNA resulting in a STOP codon (TGA, TAA or TAG).
Frameshift	S_Fs	100	Exonic insertion/deletion of a non-multiple of 3bp resulting often in a premature stop in the reading frame of the gene.
Essential splice site	S_EssentialSplice[1]	90, 95	Mutation in one of the canonical splice sites resulting in a significant effect on splicing (at least 2 out of the 3 programs indicate a relative variation in their score compared to the wild type sequence).
Start loss	S_StartLoss[1]	80, 85	Mutation leading to the loss of the initiation codon (Met).
Stop loss	S_StopLoss[1]	80, 85	Mutation leading to the loss of the STOP codon.
Intron-exon boundary	S_CloseSplice[1]	70, 75	Mutation outside of the canonical splice sites (donor site is -3 to +6', acceptor site -12 to +2) resulting in a significant effect on splicing (at least 2 out of the 3 programs indicate a relative variation in their score compared to the wild type sequence).
Missense	S_Missense[1,2]	50, 55, 60, 65, 70, 75	A single-base substitution in DNA not resulting in a change in the amino acid.
Indel in-frame	S_Inframe	40	Exonic insertion/deletion of a multiple of 3bp.
Deep intron-exon boundary	S_DeepSplice[1]	25, 30	Intronic mutation resulting in a significant effect on splicing (at least 2 out of the 3 programs indicate a relative variation in their score compared to the wild type sequence).
Synonymous coding	S_Synonymous[1]	10, 15	A single-base substitution in DNA not resulting in a change in the amino acid.

6. USAGE / OPTIONS

==========

To run VaRank, the default command line is the following:
$VARANK/bin/VaRank -vcfdir 'Path of the directory containing your vcf input file' >& log.log &

You can use "$VARANK/bin/VaRank -help" to show options.

OPTIONS:

-help	More information on the arguments
-vcfDir	Path of your study directory containing your vcf input file
-vcfInfo	To extract the info column from the .vcf file and insert the data in the outputfile (last columns). Range values: yes or no (default)
-rsfromvcf	To extract the rsID and validation status from the .vcf file and insert this in the outputfile. Range values: yes or no (default)
-nowebsearch	To allow or not the access to the web for downloading the fasta sequences for missed proteins in UniProt and/or RefSeq (only suitable when used with PolyPhen-2). It is to notice that the search can be very time consuming since getting sequences one by one. Range values: yes (default) or no
-Homstatus	To force the determination of the homozygous or heterozygous state of one variation. If set to yes it will use the Homcutoff value to decide. Range values: yes or no (default)
-Homcutoff	To determine the homozygous or heterozygous state of one variation. If set to some value it will force to reconsider the data provided. Range values: [0,100] default: 80 (active only if Homstatus=yes or when no status is given)
-MEScutoff	MaxEntScan cutoff, to determine the impact of the variant on splicing. Expressed as the % difference between the variant and the WT score. Range values: [-100,0], default: -15
-SSFcutoff	Splice Site Finder cutoff, to determine the impact of the variant on splicing. Expressed as the % difference between the variant and the WT score. Range values: [-100,0], default: -5
-NNScutoff	NNSplice cutoff, to determine the impact of the variant on splicing. Expressed as the % difference between the variant and the WT score. Range values: [-100,0], default: -10
-phastConsCutoff	To determine when a genomic position is conserved or not. Above the cutoff is considered as conserved. Range values: [0,1], default: 0.95
-readFilter	Minimum number of reads for the variants Range values: [0,-], default: 10
-depthFilter	Minimum depth for the variants Range values: [0,-], default: 10
-readPercentFilter	Minimum percent of variant reads for considering a variant Range values: [0,100], default: 15
-freqFilter	Filtering variants based on their MAF in the SNV databases (dbsnp and EVS) Range values: [0.0,1.0], default: 0.01

-rsFilter — Filtering variants on the SNP informations
Values: removeNonPathoRS (remove variants without "probable-pathogenic" or "pathogenic" annotation, see clinical significance field in dbSNP website. Filtering only for variants with at least 2 validations.)
none = keep all variants, no filtering on rsID
Default: removeNonPathoRS

-extann — Tab separated file containing annotation to add to the final output files. Restrictions for the format are: 1st line is a header, 1st column is the gene name
Typical use would be a gene file containing specific annotations such as transmission mode, disease, expression...

-metrics — Changing numerical values from frequencies to us or fr metrics (e.g. 0.2 or 0,2)
Range values: us (default) or fr

-DB — Changes the directory where the UniProt and Refseq files are stored (optional, only use if PPH2 is installed)
Ex: $VARANK/Databases (default)

-uniprot — Name of the UniProt sequence file (optional, only use if PPH2 is installed)
Ex: HUMAN.fasta.gz (default)

-refseq — Name of the RefSeq sequence file (optional, only use if PPH2 is installed)
Ex: human.protein.faa.gz (default)

-hgmdUser — HGMD User login (optional, only use if you have an HGMD license)

-hgmdPasswd — HGMD User password (optional, only use if you have an HGMD license)

The following options are provided to allow the user to modify the VaRank score corresponding to each category defined by the program:

-S_Known — Known mutation as annotated by HGMD and/or dbSNP (rsClinicalSignificance="pathogenic/probable-pathogenic").
Default: 110

-S_Nonsense — A single-base substitution in DNA resulting in a STOP codon (TGA, TAA or TAG).
default: 100

-S_Fs — Exonic insertion/deletion of a non-multiple of 3bp resulting often in a premature stop in the reading frame of the gene.
default: 100

-S_EssentialSplice — Mutation in one of the canonical splice sites resulting in a significant effect on splicing (at least 2 out of the 3 programs indicate a relative variation in their score compared to the wild type sequence)
default: 90

-S_StartLoss — Mutation leading to the loss of the initiation codon (Met).
default: 80

-S_StopLoss — Mutation leading to the loss of the STOP codon.
default: 80

-S_CloseSplice — Mutation outside of the canonical splice sites (donor site is -3 to +6', acceptor site -12 to +2) resulting in a significant effect on splicing (at least 2 out of the 3 programs indicate a relative variation in their score compared to the wild type sequence).
default: 70

-S_Missense — A single-base substitution in DNA not resulting in a change in the amino acid.
default: 50

-S_Inframe Exonic insertion/deletion of a multiple of 3bp.
default: 40

-S_DeepSplice Intronic mutation resulting in a significant effect on splicing (at least 2 out of the 3 programs indicate a relative variation in their score compared to the wild type sequence).
default: 25

-S_Synonymous A single-base substitution in DNA not resulting in a change in the amino acid.
default: 10

7. Annotations columns available in the output files

Column name	Annotation
VariantID	Variant identifier [#chr]_[genomicposition]_[RefBase]_[VarBase]
Gene	Gene symbol
omimId	OMIM® id
TranscriptID	RefSeq transcript id
TranscriptLength	Length of transcript (full cDNA length)
Chr	Chromosome of variant
Start	Start position of variant
End	End position of variant
Ref	Nucleotide sequence in the reference genome
Mut	Alternate nucleotide sequence
Uniprot	Uniprot
protein	Protein id (NCBI)
posAA	Amino acid position
wtAA_1	Reference codon
varAA_1	Alternate codon
Phred_QUAL	**QUAL:** The Phred scaled probability that a REF/ALT polymorphism exists at this site given sequencing data. Because the Phred scale is -10 * log(1-p), a value of 10 indicates a 1 in 10 chance of error, while a 100 indicates a 1 in 10^10 chance. These values can grow very large when a large amount of NGS data is used for variant calling.
HomHet	Homozygote or heterozygote status
TotalReadDepth	Total number of reads covering the position
VarReadDepth	Number of reads supporting the variant
%Reads_variation	Percent of reads supporting variant over those supporting reference sequence/base
VarType	Variant Type (substitution, deletion, insertion, duplication, delins)
CodingEffect	Variant Coding effect (synonymous, missense, nonsense, in-frame, frameshift, start loss, stop loss)
VarLocation	Variant location (upstream, 5'UTR, exon, intron, 3'UTR, downstream)
Exon	Exon (nearest exon if intronic variant)
Intron	Intron
gNomen	Genomic-level nomenclature
cNomen	CDNA-level nomenclature
pNomen	Protein-level nomenclature
rsID	dbSNP variation

rsValidation	dbSNP validated status (yes/no?
rsClinicalSignificance	dbSNP variation clinical significance
rsAncestralAllele	dbSNP ancestral allele
rsHeterozygosity	dbSNP variation average heterozygosity
rsMAF	dbSNP variation global Minor Allele
rsMAFAllele	dbSNP variation global minor allele
rsMAFCount	dbSNP variation sample size
espRefEACount	ESP reference allele count in European American population
espRefAACount	ESP reference allele count in African American population
espRefAllCount	ESP reference allele count in all population
espAltEACount	ESP alternate allele count in European American population
espAltAACount	ESP alternate allele count in African American population
espAltAllCount	ESP alternate allele count in all population
espEAMAF	Minor allele frequency in European American population
espAAMAF	Minor allele frequency in African American population
espAllMAF	Minor allele frequency in all population
espAvgReadDepth	Average sample read Depth
delta MESscore (%)	% difference between the splice score of variant with the score of the reference base
wtMEScore	WT seq. MaxEntScan score
varMEScore	Variant seq. MaxEntScan score
delta SSFscore (%)	% difference between the splice score of variant with the score of the reference base
wtSSFScore	WT seq. SpliceSiteFinder score
varSSFScore	Variant seq. SpliceSiteFinder score
delta NNSscore (%)	% difference between the splice score of variant with the score of the reference base
wtNNSScore	WT seq. NNSPLICE score
varNNSScore	Variant seq. NNSPLICE score
DistNearestSS	Distance to Nearest splice site
NearestSS	Nearest splice site
localSpliceEffect	Splicing effect in variation vicinity (New donor Site, New Acceptor Site, Cryptic Donor Strongly Activated, Cryptic Donor Weakly Activated, Cryptic Acceptor Strongly Activated, Cryptic Acceptor Weakly Activated)
SiftPred	SIFT prediction
SiftWeight	SIFT weight
SiftMedian	SIFT median
PPH2pred	PolyPhen-2 prediction (qualitative ternary classification at 10%/20% (20%/40% for HumVar) FPR thresholds (“benign”, “possibly damaging”, “probably damaging”)
phyloP	phyloP
PhastCons	PhastCons score
GranthamDist	Grantham distance
VaRank_VarScore	Prioritization score according to VaRank
AlamutAnalysis	Yes or No indicates the annotation regarding Alamut Batch analysis
Avg_TotalDepth	Total read depth average at the variant position for all samples analyzed that have the variation
SD_TotalDepth	Standard deviation associated with Avg_TotalDepth

Count_TotalDepth	Number of samples considered for the average total read depth
familyBarcode	Homozygote or heterozygote status for the sample of interest and its associted samples
Barcode	Homozygote or heterozygote status for all sample analyzed together (Hom: 2 ; Het: 1; Sample name is given at the first line of the file: ## Barcode)
Hom_Count	Number of homozygote over all samples analyzed together
Het_Count	Number of heterozygote over all samples analyzed together
Allele_Count	Number of alleles supporting the variant
Sample_Count	Total number of samples

8. FAQ

Q: What are the WARNINGs that VaRank mention while running?
A: VaRank writes to the standard output progress of the analysis including warnings about issues or missing information that can be either blocking or simply informative. More specifically while loading the VCF file(s) specific information are under survey such as vcf format consistency, patient redundancy, the total and variant read depth, the genotype, the indels. Any surveyed default will be reported to the user.

Q: I want to run a VaRank analysis again, what shall I do?
A: Simply remove all output files (*.tsv) and type the new command line. All annotations will be kept and the analysis should be done very quickly.

Q: I have already computed 5 samples in my analysis and I want to add 10 more, what should I do?
A: Considering no updated version of any annotation source or VaRank available, you can simply add the new vcf files to the already computed ones, remove all output files (*.tsv), remove simply the /Alamut/AlamutInputFile_all.txt (that will be recreated with the new variants if any) (and PPH2 input file if PPH2 is installed) and rerun VaRank. VaRank will only recompute the missing annotations and will save you the computation time of reannotating multiple times the same variants.

Q: How are the variant homozygous or heterozygous status reported?
A: VaRank trust by default the zygosity status provided by the vcf and report this in the column "Zigosity" in the output files. Nonetheless, in the case when no data is provided but total and variant depth of coverage is available, VaRank recompute this by applying the simple rule everything >=Homcutoff (default 80% see options) is homozygous and the rest is heterozygous. In order to clearly show difference with other variants those recomputed will be noted "hom?" or "het?". The same rule is applied when using the option "-Homstatus" except that variant are noted "hom" or "het".

Q: In the output files some values are set to “NA”?
A: When for a specific type of annotation no information is available then the empty value is set to "NA" (e.g. Not Available). Exception is made for several numerical columns (including rsMAF, espEAMAF, espAAMAF, espAllMAF) where "-1" is used that allows the user to further filter information without losing data.