SNVBox
Example workflow with SNVGet

After installing SNVBox

To use SNVGet:

SNVGet will retrieve feature values for a list of mutations. SNVGet therefore expects a file with a list of feature names (1 per line) as well as the list of mutations for which to retrieve features. SNVGet also requires the user to specify a name for the output file.

SNVGet commandline arguments:

-f specify file with list of features to retrieve
-c include class labels – specify separate mutation lists for each class
-o output file name
-r return raw SNVBox feature values. By default values are scaled by mean and rms.
-m allow missing values (specified by 'NA'). By default values are filled with mean.

Formatting your list of mutations:

Mutations can be specified as tab delimited lists of transcript and amino acid substitution with codon number ('transcript coordinates') or as genomic location and nucleotide substitution ('genomic coordinates'). The expected format for each coordinate system is described here:

Transcript coordinates:

- (optional) mutation identifier
- transcript
- amino acid substitution (e.g. G12D or H1047R)

Example of transcript coordinates:

```
# UID / Transcript / AA change
TR1   NM_001126116.1   D127Y
TR2   NM_001144919.1   R162Q
TR3   NM_000321.2      Q702K
TR4   NM_000245.2      A1108S
TR5   NM_004333.4      V600E
TR6   NM_001005862.1   G746V
```

Genomic Coordinates:

- (optional) mutation identifier
- chromosome (e.g. "chr1")
- 0-based start position of the nucleotide
- 1-based end position of the nucleotide

- strand of the nucleotides being reported in the following column*
- reference nucleotide
- alternative nucleotide

*If the reference and alternative nucleotides match the forward strand of the reference genome this should be “+”, if they match the reverse strand it should be “-"

Example of genomic coordinates:

```
# UID / Chr. / Start / Stop / Strand / Ref. base / Alt. base
TR1     chr17   7577505         7577506         -       G       T
TR2     chr10   123279679       123279680       -       G       A
TR3     chr13   49033966        49033967        +       C       A
TR4     chr7    116417504       116417505       +       G       T
TR5     chr7    140453135       140453136       -       T       A
TR6     chr17   37880997        37880998        +       G       T
```

Retrieving features with SNVGet:

If mutations are in transcript coordinates, SNVGetTranscript should be run to retrieve features:

```
./snvGetTranscript –f featurelist –o outputfilename.arff mutationlist
```

or if mutations are in genomic coordinates:

```
./snvGetGenomic –f featurelist –o outputfilename.arff mutationlist
```

The –c option allows the user to specify class labels to be included in the output file. The user needs to specify a separate file of mutations for each class.

```
./snvGetTranscript –c –f featurelist –o outputfilename.arff class1label mutationlist1 class2label
    mutationlist2
```

SNVGet will generate an arff file with the list of features. Please note that the arff file generated is not quite a correctly formatted arff. This is because SNVGet was intended to run with the PARF (PArallel Random Forsest) software that requires a special variant of the arff format. See http://code.google.com/p/parf/wiki/DatasetFileFormat for details.