# Package ‘sampleContamination’

February 8, 2018

**Title** Sample Contamination R Package

**Version** 0.1.0

**Description** Sample contamination discovery using variant allele frequency (VAF)

**Encoding** UTF-8

**Depends** R (>= 3.4)

**Date** 2017-08-11

**Author** Lai Ping Wong

**Maintainer** The package maintainer `<laiping.wong@roswellpark.org>`

**License** Roswell Park Cancer Institute

**LazyData** true

**RoxygenNote** 6.0.1

**Suggests** knitr,
rmarkdown

**VignetteBuilder** knitr

## R topics documented:

## filterByCov *Filter mutations base on coverage*

### Description

This function filters mutation base on coverage

### Usage

```
filterByCov(VAFdata,VAFcov,minCov)
```

### Arguments

| | |
|---|---|
| VAFdata | data frame, mutation VAF |
| VAFcov | data frame, mutation coverage |
| minCov | numeric, minimum coverage (default: 50) |

### Value

data frame, mutation VAF with coverage > minCov

### References

TBA

## filterCommonSNPs *Filter mutations base on coefficient of variant*

### Description

This function filters common SNPs base on cutoff on coefficient of variant (COV)

### Usage

```
filterCommonSNPs(VAFdata, percentage, output_path)
```

### Arguments

| | |
|---|---|
| VAFdata | data frame, mutation VAF |
| percentage | numeric, COV cutoff in percentage (default: 5) |
| output_path | character, tmp directory to store log file (SNPcv.txt) |

### Value

integer vector, row index of VAF data to be used

## References

TBA

## filterMultiSources *Filter multisource contamination*

### Description

This function filter multisource contamination

### Usage

```
filterMultiSources(filename,output_path,VAFdata,VAFcov,uniq_both)
```

### Arguments

| | |
|---|---|
| `filename` | character, contamination file name |
| `output_path` | character, output directory |
| `VAFdata` | data frame, mutation VAF |
| `VAFcov` | data frame, mutation coverage |
| `uniq_both` | integer, flag to control type of overlapping mutation in source and target samples |

### Value

write to contPredict.txt at output_path

### References

TBA

## inputData *Read in data from a file*

### Description

This function reads in file with header

### Usage

```
inputData(filename)
```

### Arguments

| | |
|---|---|
| `filename` | character, file name and its path |

### Value

data frame

### References

TBA

## `is_scalar_character` *Check for scalar argument*

### Description

This function checks for user's input is a scalar

### Usage

```
is_scalar_character(x)
```

### Arguments

| | |
|---|---|
| `x` | character (1) |

### Details

user's entry must meet the following criteria: character, length 1, not NA, non-zero length

### Value

boolean TRUE or FALSE

## `mixingRatio` *Calculate contamination level*

### Description

This function calculates contamination level, ratio of VAF_target to VAF_source

### Usage

```
mixingRatio(VAFdata,VAFcov,sample_pairs,final_rel,VAF_cutoff,VAF_ignore,
ALL_flag,output_path,sameSubject)
```

### Arguments

`VAFdata` data frame, mutation VAF

`VAFcov` data frame, mutation coverage

`sample_pairs` data frame, pairwise samples information

`final_rel` data frame, relation information

`VAF_cutoff` numeric, minimum VAF (default: 0.002)

`VAF_ignore` numeric, ignore variant less than this cutoff (default: 0.2)

`ALL_flag` logical, TRUE: calculate mixing ratio for all pairwise samples, FALSE: calculate mixing ratio for contamination pairwise samples

`output_path` character, output directory

`sameSubject` logical, TRUE: calculate contamination level for same subject pairs (default: FALSE)

### Value

matrix containing source sample, target sample, relation, contamination level, flip_flag

### References

TBA

## `multipleSource` *Identify true source from multisource contamination*

### Description

This function uses Fisher exact test to filter multiple sources that contaminate a particular target sample

### Usage

```
multipleSource(sample_pairs,VAFdata,VAFcov,VAF_cutoff,VAF_cutoff1,p.val_cutoff,output_path)
```

### Arguments

`sample_pairs` data frame, sample infomation

`VAFdata` data frame, mutation VAF

`VAFcov` data frame, mutation coverage

`VAF_cutoff` numeric, minimum VAF (default: 0.002)

`VAF_cutoff1` numeric, minimum VAF to be considered a source mutation (default: 0.05)

`p.val_cutoff` numeric, pval cutoff to be considered a significant contamination source (default: 0.05)

`output_path` character, output directory

## Value

list, remaining paiwise sample relationship information after Fisher exact test that eliminates insignificant contamination pair

## References

TBA

---

`numMutationperSample` *Count number of SNPs in each sample*

---

## Description

This function counts number of SNPs in each sample

## Usage

```
numMutationperSample (VAFdata,VAF_cutoff,n)
```

## Arguments

| | |
|---|---|
| VAFdata | data frame, mutation VAF |
| VAF_cutoff | numeric, minimum VAF |
| n | numeric, number of sample |

## Value

list, SNP count per sample

## References

TBA

---

`pairCountPoint` *Count number of SNPs in 7 regions of a given pair of sample*

---

## Description

This function counts number of SNPs in 7 regions of a given pair of sample

## Usage

```
pairCountPoint(VAFdata,pid1,pid2,slope1,slope2)
```

### Arguments

| | |
|---|---|
| `VAFdata` | data frame, mutation VAF |
| `pid1` | character, first sample id |
| `pid2` | character, second sample id |
| `slope1` | numeric, first slope that is closer to axis-X (default: 0.5) |
| `slope2` | numeric, second slope that is closer to axis-Y (default: 2) |

### Value

matrix, containing number of SNPs in 7 regions

### References

TBA

---

## `pairPCommon` *Pairwise Sample Function*

---

### Description

4 different functions for pairwise sample

### Usage

```
pairPCommon(VAFdata,VAF_cutoff,num_round_digit,n)
pairShare(VAFdata,VAF_cutoff,n)
pairList(VAFdata,n,del)
pairCommList(SNPcomm,n)
```

### Arguments

| | |
|---|---|
| `VAFdata` | data frame, mutation VAF |
| `VAF_cutoff` | numeric, minimum VAF |
| `num_round_digit` | numeric, number of rounding decimal digit (default: 3) |
| `n` | numeric, number of sample |
| `SNPcomm` | matrix, output from the calling of pairPCommon(VAFdata,VAF_cutoff,num_round_digit,n) |
| `del` | character, file delimiter |

### Details

pairPCommon: calculates pcommon

pairShare: counts number of common SNPs

pairList: generates all pairwise sample list

pairCommList: tabulates pcommon for all pairwise samples

### Value

pairPCommon: matrix, nxn pcommon, n: number of sample

pairShare: matrix, nxn number of common snps

pairList: vector, all pairwise samples

pairCommList: list, all pairwise samples and their pcommon values

### References

TBA

## pairRelation *Determine relationship between pairwise sample*

### Description

This function identifies relationship of pairwise sample [00: same patient | 10: X contaminate Y | 01 Y contaminate X | 11: both way contamination]

### Usage

```
pairRelation(sample_pairs,center_cutoff,source_cutoff,target_cutoff,
localPcomm_cutoff,region_cutoff,num_round_digit,output_path)
```

### Arguments

| | |
|---|---|
| `sample_pairs` | data frame, pairwise samples information |
| `center_cutoff` | numeric, cutoff to be classified as same subject samples (default: 0.5) |
| `source_cutoff` | numeric, cutoff to be classififed as source sample (default: 0.4) |
| `target_cutoff` | numeric, cutoff to be classified as target sample (default: 0.1) |
| `localPcomm_cutoff` | numeric, cutoff to manage high or low SNPs sharing between a pair of samples |
| `region_cutoff` | numeric, cutoff for case with low SNP sharing (default: 0.75) |
| `num_round_digit` | integer, rounding numeric up to this decimal point (default: 3) |
| `output_path` | character, output directory |

### Value

matrix, 2 columns containing pcommon and relation

### References

TBA

## `regionCountMutation` *Count number of mutation in 7 regions of VAF scatter plot for a pair of samples (s1,s2)*

### Description

This function counts number of SNPs in 7 regions of pairwise samples VAF scatter plot

### Usage

```
regionCountMutation(sample_pairs,VAFdata,SNPcount,SNPshare,VAF_cutoff,VAF_ignore,n)
```

### Arguments

| | |
|---|---|
| `sample_pairs` | data frame, pairwise samples information |
| `VAFdata` | data frame, mutation VAF |
| `SNPcount` | integer matrix, number of SNPs for each sample |
| `SNPshare` | integer matrix, number of common SNPs |
| `VAF_cutoff` | numeric, minimum VAF (default: 0.002) |
| `VAF_ignore` | numeric, ignore variant below this cutoff (default: 0.2) |
| `n` | integer, number of sample |

### Value

matrix, (n mutation x 10) containing number of SNPs in 7 regions, number of SNPs in s1, number of SNPs in s2, common SNPs between s1 and s2

### References

TBA

## `run_sampleContamination` *Main function to run sample contamination analysis*

### Description

This function detects sample contamination base on variant allele frequency (VAF) and variant coverage

### Usage

```
run_sampleContamination(data_path, output_path, config_file="config.txt",
rmSNPcv_cutoff=0,manualsetPara=FALSE,manual_localPcomm=10,manual_center=50,filterCOV=0)
```

### Arguments

`data_path` character, mutation VAF and coverage file directory

`output_path` character, output directory

`config_file` character, configuration file

`rmSNPcv_cutoff` numeric, TRUE: filter SNPs with low covariance of coefficient (COV) (default: FALSE)

`manualsetPara` logical, TRUE: manual setting base on distribution of pairwise samples commonality (default: FALSE)

`manual_localPcomm`
numeric, manual setting of localPcomm for low or high contamination (default: 10)

`manual_center` numeric, manual setting of center cutoff for same subject determination (default: 50)

`filterCOV` numeric, filter mutation below this cutoff

### Value

list, containing pcomm, center cutoff, target cutoff, source cutoff, region cutoff and number of sample

### References

TBA

## `setParameterConfig` *Set parameters base on configuration file*

### Description

This function takes in data from configuration file to set parameters

### Usage

```
setParameterConfig(configFile)
```

### Arguments

`configFile` character, configuration file name and its path

### References

TBA

# Index