Package 'OncoPhase'

August 21, 2016

Type Package

Title SOMATIC MUTATION CELLULAR PREVALENCE COMPUTATION

Version 0.1

Date 2016-02-25

Description This package offers a direct method to accurately quantify the cellular prevalence of somatic mutations in cancer using phase information. The method utilizes three sources of information: the phasing information, the copy number variation, and the allele counts. The method is demonstrated to bring more capabilities in Cancer Genomic.

LazyData TRUE

License GPL-2

Imports limSolve

Suggests

#VignetteBuilder knitr

RoxygenNote 5.0.1

NeedsCompilation no

Author Donatien Chedom-Fotso [aut, cre],
Ahmed Ahmed [aut],
Christopher Yau [aut]

Maintainer Donatien Chedom-Fotso <donatien.chedom@well.ox.ac.uk>

R topics documented:

a-OncoPhase *OncoPhase: An R package for somatic mutations cellular prevalence quantification using haplotype phasing.*

Description

OncoPhase uses haplotype phase information to accurately compute mutational cellular prevalence. OncoPhase utilizes three sources of information: the phasing information, the copy number variation, and the allele counts. It takes as input a combination of phased SNV and SNP allele-specific sequence read counts and local allele-specific copy numbers to determine the proportion of cells harboring the SNV and compute specific and detailed mutation cellular prevalence for each of the following groups of cells:

Details

Germ Germline cells having a normal genotype with no mutations and no copy number alteration at the considered locus.

Alt Cells harboring one alternative between the two somatic alterations. That is either only the SNV if C=1 (SNV occurred before SCNA) or only the SCNA if C=0 (SNV occurred after the SCNA).

Both Cells harboring both somatic alterations. That is the SNV and the SCNA

OncoPhase can compute the mutation cellular prevalence under three different modes : PhasedSNP, FlankingSNP and SNVOnly.

The main functions for somatic mutation cellular prevalence computation are `getPrevalence` , `getSamplePrevalence` and `getMultiSamplesPrevalence` . For more detailed information on usage, see the package vignette, by typing `vignette("OncoPhase")`. All support questions should be emailed to the authors.

Author(s)

Donatien Chedom-Fotso, Ahmed Ahmed, Christopher Yau.

References

OncoPhase reference:

Chedom-Fotso Donatien, Ahmed Ashour Ahmed, and Christopher Yau. "OncoPhase: Quantification of somatic mutation cellular prevalence using phase information." bioRxiv (2016): 046631.

chr22_XYZ101 *chr22_XYZ101 : Patient XYZ101 data from chromosome 22 .*

Description

A generated dataset containing allele counts, haplotype phasing and copy number information on chromosome 22 for a patient XYZ101 (Data created) .

Usage

```
chr22_XYZ101
```

Format

Contains the following data : :

tumoursamples The list of tumor samples of the study

SNP_allelecount_df Data frame containing the count of allele supporting the variant of each mutations

Chrom : Chromosomes

Start : Starting position

End : Position of the mutation

Vartype : variant Type

IsGermline : is the mutation a Germline SNP or a Somatic mutation

Ref : Reference sequence

All : Variant sequence

One entry per tumor samples

ref_allelecount_df Data frame containing the count of allele supporting the reference at each mutation

phasing_association_df A data frame containing for each somatic mutations, a colon separated list of Germline mutations phased to it.

major_copynumber_df A data frame containing the major copy number of the mutations at each tumor samples

minor_copynumber_df A data frame containing the minor copy number of the mutations at each tumor samples

minor_copynumber_df A data frame containing the normal cell contamination rate for each mutations at each tumor samples

Details

Also available on the same patient : chr10, chr15 and chr18

`getMatrices` *Generate the matrices C, W and M from a set of parameters.*

Description

This is a generic function to generate the matrices of the linear system (see the paper) from the allele counts and the copy number information.

Usage

```
getMatrices(varcounts_snv, refcounts_snv, major_cn, minor_cn, varcounts_snp,
  refcounts_snp, context, LocusCoverage = FALSE)
```

Arguments

varcounts_snv	A count of alleles supporting the variant sequence of the somatic mutation
refcounts_snv	A count of alleles supporting the reference sequence of the somatic mutation
major_cn	major copy number at the locus of the mutation
minor_cn	minor copy number at the locus of the mutation
varcounts_snp	A count of alleles supporting the variant sequence of the Germline SNP
refcounts_snp	A count of alleles supporting the reference sequence of the Germline SNP
context	represents either the situation of a mutation which occurred after the CNV ("C1") or the context of a mutation which occurred before the CNV ("C2"). If not provided, the right context will be estimated from the input
LocusCoverage	when set to TRUE, the SNV locus coverage is estimated to the average coverage of the phased SNP and the variant allele fraction is the ratio of the variant allele count over the estimated locus coverage.

Value

the matrices W, C and M for the linear system of prevalence computation.

Examples

```
Matrices = getMatrices(3, 10,2,1,8,5,"C1")

print(Matrices)
#$context
#[1] "C1"
#
#$W
#          SNP       SNV
#SNP 0.6153846 0.0000000
#SNV 0.0000000 0.2307692
#
#$M
#    Germ Alt Both
#SNP    1   2    2
#SNV    0   0    1
#
#$C
#    Germ Alt Both
#SNP    2   3    3
#SNV    2   3    3
```

---

getMatricesSNVOnly *Generate the matrices C, W and M from a set of parameters under the mode "SNVOnly.*

Description

This is a generic function to generate the matrices of the linear system (see the paper) from the allele counts and the copy number information under the SNVOnly mode.

Usage

```
getMatricesSNVOnly(varcounts_snv, refcounts_snv, major_cn, minor_cn, context,
  sigma = NULL)
```

Arguments

varcounts_snv	A count of alleles supporting the variant sequence of the somatic mutation
refcounts_snv	A count of alleles supporting the reference sequence of the somatic mutation
major_cn	major Copy number at the locus of the mutation
minor_cn	minor copy number at the locus of the mutation
context	represents either the situation of a mutation which occurred after the CNV ("C1") or the context of a mutation which occurred before the CNV ("C2"). If not provided, the right context will be estimated from the input
sigma	Copy number of the parental chromosome harboring the mutation.

Value

the matrices W, C and M for the linear system of prevalence computation.

Examples

```
Matrices = getMatricesSNVOnly(3,10,2,1,"C1")

print(Matrices)
#$context
#[1] "C1"
#
#$W
#SNV
#SNV 0.2307692
#
#$M
#Germ Alt Both
#SNV    0   0    1
#
#$C
#Germ Alt Both
#SNV    2   3    3
```

`getMultiSamplesPrevalence` *Somatic mutations cellular prevalence computation using haplotype phasing on a multiple sample study.*

Description

This is a generic function to compute the cellular prevalence of somatic mutations in cancer using haplotype phasing. The function applies the model to a range of mutations located at a given genomic region or at the whole genome scale. The model computes the prevalence of a somatic mutation relatively to close and eventually phased germline mutations. It uses three sources of information as input : The allelic counts, the phasing information and the copy number alteration. Multiple tumor samples can be provided for the prevalence computation.

Usage

```
getMultiSamplesPrevalence(snp_allelecount_df, ref_allelecount_df,
  major_copynumber_df, minor_copynumber_df, mode = "PhasedSNP",
  phasing_association_df = NULL, NormalCellContamination = NULL,
  nbFirstColumns = 3, tumoursamples = NULL, region = NULL,
  detail = TRUE, LocusRadius = 10000, LocusCoverage = TRUE,
  SomaticCountAdjust = TRUE)
```

Arguments

`snp_allelecount_df`
: A data frame containing for each mutation the allelic Counts of the variant at each tumor samples. The data frame should contain at least the following three columns among its firsts columns: Chrom (The mutation chromosome) , Pos or End (The mutation position) and IsGermline (is the mutation a germline or somatic mutation).

`ref_allelecount_df`
: A data frame containing for each mutation the allelic count of the reference at each tumor sample. The data frame should contain at least the following three columns among its firsts columns: Chrom (The mutation chromosome) , Pos or End and IsGermline (is the mutation a Germline or Somatic mutation)

`major_copynumber_df`
: A data frame containing for each mutation, its major chromosomal copy number at each tumor samples. Should contain at least the following three columns among its firsts columns: Chrom (The mutation chromosome) , Pos or End (The mutation position) and IsGermline (is the mutation a Germline or Somatic mutation)

`minor_copynumber_df`
: A data frame containing for each mutation the minor chromosomal copy number at each tumor samples. Should contain at least the following three columns among its firsts columns: Chrom (The mutation chromosome) , Pos or End (The mutation position) and IsGermline (is the mutation a Germline or Somatic mutation)

`mode`
: The mode under which the prevalence is computed (default : PhasedSNP , alternatives modes are FlankingSNP, OptimalSNP and SNVOnly). Can also be provided as a numeric 0=SNVOnly, 1= PhasedSNP, 2=FlankingSNP and 3 = OptimalSNP

`phasing_association_df`
: A data frame containing for each somatic mutation, a colon separated list of germline SNP phased to it.

`NormalCellContamination`
: If provided, represents the rate of normal cells contaminations in the experiment.

`nbFirstColumns` Number of first columns in snp_allelecount_df to reproduce in the output dataframe e.g: Chrom, Pos, Vartype. Columns from nbFirstColumns +1 to the last column should contains the information needed for the prevalence computation at each tumor sample

`tumoursamples` The list of tumor samples to consider for the prevalence computation. These samples should be present as column headers in the data frame snp_allelecount_df, ref_allelecount_df, major_copynumber_df,minor_copynumber_df. If not provided, the headers from nbFirstColumns + 1 to the last column of snp_allelecount_df are retrieved and their intersection with the other inputted data frames headers is considered

`region` The region of the genome to consider for the prevalence computation in the format chrom:start-end e.g "chr22:179800-98767

`detail` when set to TRUE, a detailed output is generated containing, the context and the detailed prevalence for each group of cells (germline cells, cells affected by one of the two genomic alterations (SNV or CNV) but not both, cells affected by both copy number alteration and SNV).

`LocusRadius` Only phased SNPs located within LocusRadius bp from the somatic mutation will be considered.

`LocusCoverage` when set to TRUE, the SNV locus coverage is estimated to the average coverage of the phased SNP and the variant allele fraction is the ratio of the variant allele count over the estimated locus coverage.

`SomaticCountAdjust`
when set to TRUE, varcounts_snv and refcounts_snv might be adjusted if necessary so that they meet the rules varcounts_snv <= varcounts_snp, refcounts_snv >= refcounts_snp and varcounts_snv + refcounts_snv ~ Poiss(varcounts_snp + refcounts_snp). Not used if mode=SNVOnly.

Value

A data frame containing :

Column 1 to NbFirstcolumn of the input data frame snp_allelecount_df. This will generally include the chromosome and the position of the mutation plus any other columns to report in the prevalence data frame (e.g REF, ALL, ...)

One column per tumor sample reporting the prevalence of the mutation at each samples

Examples

```
#Example 1: Loading a simple example data set with two somatic mutations, 5 germlines SNP
# and 3 tumor samples
data(simpleExample2)
se=simpleExample2
prevalence_df=getMultiSamplesPrevalence(se$snp_allelecount_df, se$ref_allelecount_df,
 se$major_copynumber_df,se$minor_copynumber_df,phasing_association_df=se$phasing_association_df, )
print(prevalence_df)

#Chrom     End IsGermline  Tumour1        Tumour2        Tumour3
#mutation2  chr2 3003000          0 C2:0|0|1 C2:0.15|0|0.85 C2:0.12|0|0.88
#mutation6  chr2 4008000          0 C1:1|0|0       C1:1|0|0 C2:0|0.24|0.76

#Example 2 : Computing somatic mutation cellular prevalence on chromosome 15 of  patient XYZ101
# (data created)
```

```
data("chr22_XYZ101")
ds=chr22_XYZ101
masterprevalence_df=getMultiSamplesPrevalence(ds$snp_allelecount_df, ds$ref_allelecount_df,
  ds$major_copynumber_df,ds$minor_copynumber_df,phasing_association_df = ds$phasing_association_df,
    nbFirstColumns=6,detail=FALSE)
print(head(masterprevalence_df))

data("chr18_XYZ101")
df=chr18_XYZ101
masterprevalence_df=getMultiSamplesPrevalence(df$snp_allelecount_df, df$ref_allelecount_df,
 df$major_copynumber_df,df$minor_copynumber_df,phasing_association_df=df$phasing_association_df,
 nbFirstColumns=6, region="chr18:10000000-80000000")
print(head(masterprevalence_df))



#'@seealso \code{\link{getPrevalence}}
```

`getPrevalence` *Computes cellular prevalence at a single mutation point*

Description

This is a generic function to compute the cellular prevalence of a somatic mutation point using OncoPhase method. The method computes the prevalence of the somatic mutation relatively to phased nearby SNPs whose prevalence are known to be 1. `getPrevalence` requires the allelic-information of the somatic mutation and the aggregated information of its Phased SNP but the function can also be run in the absence of phasing information (FlankingSNP mode) or nearby SNP (SNVOnly mode).

Usage

```
getPrevalence(varcounts_snv, refcounts_snv, major_cn, minor_cn,
  varcounts_snp = NULL, refcounts_snp = NULL, detail = 0,
  mode = "PhasedSNP", Trace = FALSE, LocusCoverage = TRUE,
  SomaticCountAdjust = TRUE, NormalCellContamination = NULL)
```

Arguments

`varcounts_snv`	A count (or a vector of counts if multiple samples) of alleles supporting the variant sequence of the somatic mutation
`refcounts_snv`	A count (or a vector of counts if multiple samples) of alleles supporting the reference sequence of the somatic mutation
`major_cn`	major copy number (or a vector if multiple samples) at the locus of the mutation
`minor_cn`	minor copy number (or a vector if multiple samples) at the locus of the mutation
`varcounts_snp`	A count (or a vector of counts if multiple samples) of alleles supporting the variant sequence of the Germline SNP
`refcounts_snp`	A count (or a vector of counts if multiple samples) of alleles supporting the reference sequence of the Germline SNP

`detail` when set to 0, the function simply output the cellular prevalence of the somatic mutation. if set to 1, a detailed output is generated containing:

Context The associated context (C=1 or C=2)

Prevalence The computed cellular prevalence

DetailedPrevvalence the detailed prevalence for each group of cells (germline cells (Germ), cells affected by one of the two genomic alterations (Alt), cells affected by both genomic alterations (Both)

ResidualNorm Norm of the residuals when an approximation is performed

CondensedPrevalence A colon separated list of the above fields (Context, Prevalence, Detailedprevalence and ResidualNorm). The detailed prevalence are separated by "|"

if detail is set to 2, the function outputs the CondensedPrevalence field above.

`mode` The mode under which the prevalence is computed (default : PhasedSNP , alternatives methods are FlankingSNP and SNVOnly). Can also be provided as a numeric 0=SNVOnly, 1= PhasedSNP, 2=FlankingSNP

`Trace` if set to TRUE, print the trace of the computation.

`LocusCoverage` when set to TRUE, the SNV locus coverage is estimated to the average coverage of the phased SNP and the variant allele fraction is the ratio of the variant allele count over the estimated locus coverage.

`SomaticCountAdjust`
when set to 1, varcounts_snv and refcounts_snv might be adjusted if necessary so that they meet the rules varcounts_snv <= varcounts_snp, refcounts_snv >= refcounts_snp and varcounts_snv + refcounts_snv ~ Poiss(varcounts_snp + refcounts_snp). Not used if mode=SNVOnly,

`NormalCellContamination`
If provided, represents the rate of normal cells contaminations in the experiment.

Details

The method particularly exhibits an increase in the accuracy when the locus of the SNP is also affected by a somatic copy number alteration (SCNA). The method detect the temporal relationship between the two alterations (C1: SNV occurred after the SCNA; C2: SNV occurred before the SCNA) and computes the detailed prevalence of each of the following group of cells (if detail is set to TRUE) :

Germ Cells having a germline genotype at the locus of the SNV. That is No SNV, no SCNA

Alt Cells having one alternative of the two somatic alteration. That is either the SCNA, either the SNV not both.

Both Cells having both somatic alterations. That is the SNV and the SCNA

OncoPhase can be run under three modes:

PhasedSNP The prevalence is computed relatively to a Phased SNP

FlankingSNP In the absence of phasing information, the prevalence is computed relatively to a neighbor SNP located on the same locus with the somatic SNV. NA is returned if the prevalence cannot be resolved without knowing the phasing information between the SNP and the SNV.

SNVOnly The prevalence is computed using only the SNV information without the usage of any nearby SNP

Value

The cellular prevalence if detail =0, a detailed output if detail = 1, and a condensed output if detail =2. See the usage of the parameter detail above.

See Also

getPrevalence, getSamplePrevalence, getSinglePrevalence, getPrevalenceSNVOnly

---

`getPrevalenceSNVOnly` *Compute the cellular prevalence of each group of cells in case of SNVOnly mode*

Description

This is a generic function to compute the detailed prevalence of a single mutation using the linear system making the model.

Usage

```
getPrevalenceSNVOnly(varcounts_snv, refcounts_snv, major_cn, minor_cn, context,
  sigma = NULL, Trace = FALSE, NormalCellContamination = NULL)
```

Arguments

`varcounts_snv`	A count of alleles supporting the variant sequence of the somatic mutation
`refcounts_snv`	A count of alleles supporting the reference sequence of the somatic mutation
`major_cn`	major copy number at the locus of the mutation
`minor_cn`	minor copy number (or a vector of copy number if multiple tumor samples)
`context`	represents either the situation of a mutation which occurred after the CNV ("C1") or the context of a mutation which occurred before the CNV ("C2"). If not provided, the right context will be estimated from the input
`sigma`	The parental copy number of the chromosome harboring the mutation locus. Only needed if the context = C2. Should be either the major copy number either minor copy number
`Trace`	If TRUE, a trace of the execution will be printed
`NormalCellContamination`	If provided, represents the rate of normal cells contaminations in the experiment.

Value

A list of the three cellular prevalence of each of the three groups of cells

See Also

getPrevalence, getMatricesSNVOnly

Examples

```
Prevalences = getPrevalenceSNVOnly(3,10,2,1,"C2",2)

print(Prevalences)
#Germ      Alt     Both residual
#0.60     0.31     0.09     0.00
```

`getSamplePrevalence` *Somatic mutations cellular prevalence on a sample.*

Description

This function computes the cellular prevalence of a list of somatic mutations in cancer. The function applies the model to a range of mutations located at a given genomic region or at the whole genome scale. The function invokes `getPrevalence` to compute the cellular prevalence for each mutation of the list. The model computes the prevalence of a somatic mutation relatively to close and eventually phased germline SNP as specified in `getPrevalence`.

Usage

```
getSamplePrevalence(input_df, mode = "PhasedSNP", nbFirstColumns = 0,
  region = NULL, detail = TRUE, LocusCoverage = TRUE,
  SomaticCountAdjust = TRUE, NormalCellContamination = NULL)
```

Arguments

`input_df`	A data frame containing for each mutations : **varcounts_snv** Alelle counts supporting the SNV **refcounts_snv** Alelle counts supporting the reference at the SNV locus **major_cn** Major copy number at the SNV locus **minor_cn** Minor copy number at the SNV locus **varcounts_snp** Alelle counts supporting the SNP **refcounts_snp** Alelle counts supporting the reference at the SNP
`mode`	The mode under which the prevalence is computed (Default : PhasedSNP , alternatives methods are FlankingSNP, OptimalSNP,and SNVOnly). Can also be provided as a numeric 0=SNVOnly, 1= PhasedSNP, 2=FlankingSNP and 3 = OptimalSNP
`nbFirstColumns`	Number of first columns in input_df to reproduce in the output dataframe e.g: Chrom, Pos, Vartype. Columns from nbFirstColumns +1 to the last column should contains the information needed for the prevalence computation.
`region`	The region of the genome to consider for the prevalence computation in the format chrom:start-end e.g "chr22:179800-98767.
`detail`	when set to TRUE, a detailed output is generated containing, the context and the detailed prevalence for each group of cells (germline cells, cells affected by one of the two genomic alterations (SNV or CNV) but not both, cells affected by both copy number alteration and SNV).

`LocusCoverage` when set to TRUE, the SNV locus coverage is estimated to the average coverage of the phased SNP and the variant allele fraction is the ratio of the variant allele count over the estimated locus coverage.

`SomaticCountAdjust`
when set to TRUE, varcounts_snv and refcounts_snv might be adjusted if necessary so that they meet the rules varcounts_snv <= varcounts_snp, refcounts_snv >= refcounts_snp and varcounts_snv + refcounts_snv ~ Poiss(varcounts_snp + refcounts_snp). Not used if mode=SNVOnly,

`NormalCellContamination`
If provided, represents the rate of normal cells contaminations in the experiment.

Value

A data frame containing :

Column 1 to NbFirstcolumn of the input data frame input_df. This will generally include the chromosome and the position of the mutation plus any other columns to report in the prevalence dataframe (e.g REF, ALL, ...)

and the following information

Prev The Cellular Prevalence of the mutation

Germ The proportion of cells with a normal genotype

Alt The proportion of cells with only the CNA if the context C=C1 or with only the SNV if the context C=C2

Both The proportion of cells with both the SNV and the SCNA

Context Context at the mutation. If C1 then the SNV occurred after the SCNA, if C=c2 then the SNV occurred before the SCNA

residual Residual after limSolve approximation.

Examples

```
#Example 1:

input_file=system.file("extdata","phylogeny1_d300_n80.tsv", package = "OncoPhase")
input_df<-read.table(input_file,header=TRUE)
rownames(input_df) = input_df$mutation_id
print(input_df)
#  mut_id varcounts_snv refcounts_snv major_cn minor_cn varcounts_snp refcounts_snp
#a      a      151  152        1        1      151  135
#b      b      123  176        1        1      161  150
#c      c       94  209        2        1      176  134
#d      d       23  283        1        1      155  144
#e      e       60  228        2        0      174  125

prevalence_df=getSamplePrevalence(input_df,nbFirstColumns = 1)

print(prevalence_df)
#  mut_id   Prev   Germ    Alt   Both Residual Context
#  a      a 0.9967 0.0017 0.0017 0.9967  3.1e-03      C1
#  b      b 0.8230 0.0890 0.0890 0.8230  1.3e-03      C1
#  c      c 0.4010 0.6000 0.0910 0.3100  3.9e-33      C2
#  d      d 0.1500 0.4200 0.4200 0.1500  1.4e-03      C1
#  e      e 0.2490 0.7500 0.0890 0.1600  5.1e-31      C2
```

```
#'@seealso \code{\link{getPrevalence}}
```

`getSinglePrevalence` *Compute the cellular prevalence of each group of cells*

Description

This is a generic function to compute the detailed prevalence of a single mutation using the linear system of the model.

Usage

```
getSinglePrevalence(varcounts_snv, refcounts_snv, major_cn, minor_cn,
  varcounts_snp, refcounts_snp, context, Trace = FALSE,
  LocusCoverage = TRUE, NormalCellContamination = NULL)
```

Arguments

`varcounts_snv`	A count of alleles supporting the variant sequence of the somatic mutation
`refcounts_snv`	A count of alleles supporting the reference sequence of the somatic mutation
`major_cn`	major copy number at the locus of the mutation
`minor_cn`	minor copy number (or a vector of copy number if multiple tumor samples)
`varcounts_snp`	A count of alleles supporting the variant sequence of the Germline SNP
`refcounts_snp`	A count of alleles supporting the reference sequence of the Germline SNP
`context`	represents either the situation of a mutation which occurred after the CNV ("C1") or the context of a mutation which occurred before the CNV ("C2"). If not provided, the right context will be estimated from the input
`Trace`	Print a trace of the eecution.
`LocusCoverage`	when set to TRUE, the SNV locus coverage is estimated to the average coverage of the phased SNP and the variant allele fraction is the ratio of the variant allele count over the estimated locus coverage.
`NormalCellContamination`	If provided, represents the rate of normal cells contaminations in the experiment.

Value

A list of the three cellular prevalence of each of the three groups of cells

See Also

`getPrevalence`, `getMatrices`

Examples

```
Prevalences = getSinglePrevalence(3, 10,2,1,8,5,"C1")

 print(Prevalences)
# Germ  Alt Both
# 0.4  0.0  0.6
```

Index