

Package ‘sCNaphase’

November 22, 2015

Type Package

Title Esitmate tumor CN profile and tumor cellularity.

Version 1.4.1

Date 2014-08-22

Author Wenhan CHEN

Maintainer Wenhan <wenhan.chen@uq.net.au>

Description Esitmate tumor CN profile and tumor cellularity.

Depends Rcpp, BH (>= 1.58.0)

License L-GPL

LinkingTo BH, Rcpp

Imports BH, Rcpp

Date/Publication 2015-Nov-22

R topics documented:

How to install	2
Package overview	2
Infer copy number	3
Output Formatting	5
Index	6

How to install

How to install the tool and address the dependencies

Description

sCNPhase is dependent on the **NLOPT**, which is a C++ library of local and global optimization algorithms. The NLOPT library is expected to be installed beforehand. To locate the header and the library files of NLOPT, the follow two environment variables needs to be set:

```
export PKG_NLOPT_LIBS="/folder/to/NLOPT/lib"
export PKG_NLOPT_INCLUDE="/folder/to/NLOPT/include"
```

sCNPhase employs the **Boost C++ library** for arithmetic calculations. This dependency can be addressed by installing the BH package from CRAN. After these two steps, the following command will install the sCNPhase package.

```
R CMD INSTALL ./sCNPhase
```

For faster speed, sCNPhase enables parallel computing using **OpenMP**, so that it can run on multiple CPUs. This installation procedure has been tested on Linux with GCC-4.4.3.

Note

This compiling procedure was tested on centos, ubuntu with GCC-4.4.3, R-3.0.1.

Package overview

sCNPhase: somatic copy number profiling based on allelic phases.

Description

sCNPhase is an R package, designed for estimating haploid somatic copy number alternations (sCNAs), tumor cellularity and tumor ploidy, based on whole genome sequencing (WGS) or whole exome sequencing (WES) data. To detect the somatic alterations and avoid the GC bias, a patient-matched normal samples is required and expected to be sequenced based on the same protocol and platform as the tumor samples.

To generate the input for sCNPhase, a pre-processing step is expected to produce two sets of vcf files and a set of haps files.

1. Set N: a vcf file with germ-line SNPs called from a normal sample for each chromosome.
2. Set T: a vcf file called from a tumor sample at the germ-line SNPs for each chromosome.
3. Set H: a hap file with the phase information for each germ-line SNP for each chromosomes.

Set N and Set T is generated from samtools mpileup.

Set C is calculated based on Set A using **SHAPEIT**.

Given these files, sCNPhase generates the sCNAs profile at haploid level, the tumor purity and ploidy estimations through the inferCNA function. The sCNAs profile can be formatted into segmentation files, vcf files or visualized in d.SKY plots.

Note

Before running the following code in command line, specify the the number of CPUs by

```
export OMP_NUM_THREADS=2
```

so that this would run on 2 CPUs.

Author(s)

Wenhan CHEN

ReferencesWenhan CHEN, sCNPhase package, <https://github.com/Yves-CHEN/sCNPhase>**See Also**[inferCNA](#) [genSegFile](#) [produceDSKY](#)**Examples**

```
# -----
# This is a demo written in R, with minimal number of parameters.
# -----
anaName = "inferCN" # This specifies the name of the analysis.
# It can be any name, but better to be meaningful and unique,
# as the result files are made up of anaName.

chroms = c(1:22) # This specify chromosomes of genome to consider.
# 1:22 means the 22 autosomes.

nPrefix = "/baseDir1/filename" # inferCNA function will try to locate
# the a vcf for a normal genome,
# named as /baseDir1/filename.chr{1...22}.vcf, (Set N)
# a hap file /baseDir1/filename.chr{1...22}.haps (Set H)

tPrefix = "/baseDir2/filename" # inferCNA will try to locate the a vcf
# for a tumor genome, named as /baseDir2/filename.chr{1...22}.vcf (Set T)

inferCNA (anaName, nPrefix, tPrefix, chroms) # inferCNA will profile
# sCNAs and generate a R dat file called res.{anaName}.phased.chr.W.dat
# in the current directory, based on which sCNPhase can then format
# the estimation to the segmentation file, the d.SKY plot and a vcf file.

genSegFile(anaList= anaName, outdir = "test") # This generates a *.csv file.
# Each row corresponds to a particular sCNAs with chrID, start, end,
# copy number, allelic copy number.

produceDSKY(anaList= anaName, outDir = "test")
# This will generate the d.SKY plot into a pdf file.
```

Infer copy number *A function for haplotype-based allelic copy number alteration inference.*

Description

This function performs haplotype-based allelic copy number alteration inference on the paired normal-tumor sequencing data. The preprocessing is needed to call the variances and produce the haplotypes. This function has no return values, but instead produces a *.dat file containing all the information about the estimation. This *.dat file can be further processed into a vcf files, segmentation file or a d.SKY plot.

The vcf files and the file with the phase information has to be named in the format of \$tPrefix.chr*.vcf and \$Prefix.chr*.haps and placed in the same folder.

Usage

```
inferCNA <- function(anaName, nPrefix, tPrefix, chroms, doPhase = T,
  forceRead=F, maxCopyNum=12, mlen = 30, maxiter = 1,
  ploidy = seq(1, 2.5, 0.1), allelicMapability = F, generateLog = T)
```

Arguments

anaName	Any label for the analysis. The anaName appears in the name of output files.
nPrefix	The prefix of the normal sample files in vcf format. The prefix can include the path specification to the vcf files. For example, nPrefix = "../Data/HCC1143.normal"
tPrefix	The prefix of the tumor sample files in vcf format. Same as above.
chroms	The chromosomes to be included in analysis. By default, chroms=1:22 which represents all the autosomes.
doPhase	If this value is TRUE, the phase information will be included. Default value is T. The analysis will be based on haplotypes.
forceRead	If this value is F, the function will load the depths information from a temporary *.dat file. If the *.dat doesn't exist, the function will read the depth information from the vcf files specified by nPrefix and tPrefix and generate a temporary *.dat file. Default value is F, since parsing vcf file can cost time. If it is T, the function will disregard these temporary *.dat file
maxCopyNum	The upper limit of the somatic copy number alteration from normal. Default value is 12. The lower limit is 0 for genotype of homozygous deletions.
mlen	By default, the sCNPhase performs a merge on the SNPs. This specifies how many allelic depths will be merged. Default value is 30 for every 30 SNPs. The greater improves the power of the model for estimation, but reduces the resolution.
maxiter	The function performs an EM estimation. This value specifies the number of iterations. Default value is 1.
ploidy	This specifies a list of possible ploidy values to search, since the average ploidy for cancer is often unknown. sCNPhase chooses the value that maximize the likelihood function. The value of 1 corresponds to ploidy index of 1, which is ploidy of 2. The default values is 1 to 2.5 incremented by 0.1. However when a rough value of the average is available, this prior knowledge can be fed to sCNPhase through this parameter.
allelicMapability	Allelic bias can cause the imbalance of the allelic depth, even when the copy number of reference allele equals to the alternative allele. When the allelicMapability is T, sCNPhase will adjust this bias by correcting the number of reads mapped to each allele.

`generateLog` If this value is TRUE, sCNaphase will generate a pdf report about the analysis. The default value is T.

Note

Many of the parameters provided can just keep the default value. The `anaName`, `nPrefix`, `tPrefix` are necessary, and required to be specified each time. The `ploidy` parameter is useful, when a rough value of the average ploidy is known.

Output Formating *Generate CN segmentation, vcf files and d.SKY plot*

Description

By default, sCNaphase generates a R data file `*.dat` which contains all the information. This raw result can be formatted to more readable outputs using `genSegFile`, `produceDSKY`.

The `genSegFile` produces a `*.csv` file, similar to `*.bed` file format, which includes 5 columns:

```
chr  start  end  CN  mCN
```

which stands for chromosome identifier, start position, end position, the overall copy number, the copy number for the less amplified allele respectively. If cases of LOHs and heterozygous deletions, `mCN` is 0.

The digital SKY plot looks like spectral karyotyping (SKY) plots, which shows the chromosomes in cytobands and marks the copy number for each regions, the regions with LOHs, the regions with heterozygous deletions. An example of this can be found at ([url N.A.](#)). This allows visual inspections of chromosomal, focal copy number deletions, gains or amplification. Generation of the cytobands requires the `quantsmooth` R packages.

The function `produceCytoPlot` generates `cytoPlot`, which is a 2D plot with a cytoband on the x-axis, and copy number on the y-axis. The function `vcfFile` outputs the copy number, allelic phase information to the existing `vcf` file. The requires the `VariantAnnotation` package from Bioconductor.

Usage

```
genSegFile (anaList= anaName, outdir = "test")
genVCFFile (anaName, vcfFile, chrID, outDir = "./")
produceCytoPlot (anaList = anaName, col = 2, outDir = "./")
produceDSKY(anaList= anaName, outdir = "test")
```

Arguments

<code>anaName</code>	Any label for the analysis. This should be the same as the <code>anaName</code> specified for function <code>inferCNA</code>
<code>outdir</code>	The output directory.
<code>vcfFile</code>	<code>vcfFile</code> corresponds to an existing <code>vcfFile</code> , into which sCNaphase output the copy number formation.
<code>col</code>	<code>col</code> stands for color. Be default, the output color is 2 (red).
<code>chrID</code>	The chromosome identifier without the <code>chr</code> prefix, e.g. 1, X, for <code>chr1</code> and <code>chrX</code> .

Examples

```
genSegFile(anaList= anaName, outdir = "test")    # This generates a *.csv file.  
# Each row corresponds to a particular sCNAs with chrID, start, end,  
# copy number, allelic copy number.  
  
produceDSKY(anaList= anaName, outDir = "test")  
# This will generate the d.SKY plot into a pdf file.
```

Index

- *Topic **allelic phases**
 - Package overview, [2](#)
- *Topic **copy number**
 - Package overview, [2](#)
- *Topic **haploid copy number**
 - Package overview, [2](#)
- *Topic **somatic**
 - Package overview, [2](#)
- *Topic **tumor cellularity**
 - Package overview, [2](#)
- *Topic **tumor ploidy**
 - Package overview, [2](#)
- *Topic **tumor purity**
 - Package overview, [2](#)

genSegFile, [3](#)

genSegFile, produceDSKY (*Output Formating*), [5](#)

How to install, [1](#)

Infer copy number, [3](#)

inferCNA, [3](#)

inferCNA (*Infer copy number*), [3](#)

inferCNA (*Output Formating*), [5](#)

Output Formating, [5](#)

Package overview, [2](#)

produceDSKY, [3](#)

scNAPhase, scNAPhase (*How to install*), [1](#)

scNAPhase, scNAPhase, scNAPhase, scNAPhase-package (*Package overview*), [2](#)