Package ‘sCNAphase’

February 15, 2016

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` *How to install the tool and address the dependencies*

Description

sCNAphase is dependent on the NLOPT[1], which is a C++ library of local and global optimation 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"
```

sCNAphase employs the Boost C++ library[2] for arithmetic calculations. This dependency can be addressed by installing the BH package from CRAN. After these two steps, the following command

[1] http://ab-initio.mit.edu/wiki/index.php/NLopt_C-plus-plus_Reference

[2] http://www.boost.org/

will install the sCNAphase package.

```
R CMD INSTALL ./sCNAphase
```

For faster speed, sCNAphase enables parallel computing using OpenMP[3], 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.

`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 analaysi. By defaut, 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.

[3] `http://openmp.org/wp/`

`mlen` By default, the sCNAphase 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. sCNAphase chooses the value that maximize the likelihoood 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 sCNAphase 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, sCNAphase 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 neccessary, 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 chromsome 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 spectural karyotypiing (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 (https://figshare.com/authors/_/1365237). This allows visual inspections of chromsomal, focal copy number deletions, gains or amplification. Generation of the cytobands requires the quantsmooth R packages.

Usage

```
genSegFile(anaList= anaName, outdir = "test",
                        mlRemoveLevel = 0.04, ifload = T)
produceDSKY(anaList= anaName, outdir = "test",
              mlRemoveLevel = 0.04, ifload = T)
```

Arguments

`anaName` Any label for the analysis. This should be the same as the anaName specified for function inferCNA

`outdir` The output directory.

`mlRemoveLevel` When merging phased allelic depths into PHF, the merging error can rise at the boundary between two neighbor sCNAs. mlRemoveLevel determines to which degree, the PHFs needs to be filtered due to merging error. By default, the value is 0.04, so that 4% of the PHFs will be removed.

`ifload` By default, the anaName allows the function to identify the location of a *.dat. Then within the functions, the *.dat file is loaded. However, if the ifload is set to F, the *.dat file can be specified and loaded from outside this function.

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.
```

`Package overview` *sCNAphase: somatic copy number profiling based on allelic phases.*

Description

sCNAphase 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 protocal and platform as the tumor samples.

To generate the input for sCNAphase, 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 chromosome.

Set N and Set T is generated from samtools mpileup.
Set C is calculated based on Set A using SHAPEIT[4].

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

[4] https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html

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

References

Wenhan CHEN, sCNAphase package, `https://github.com/Yves-CHEN/sCNAphase`

See Also

`inferCNA` `genSegFile` `produceDSKY`

Examples

```
# ------------------------------------------------------------------------------
# This is a demo written in R, with minimal number of paramenters.
# ------------------------------------------------------------------------------
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 sCNAphase 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.
```