

# Package ‘TitanCNA’

April 12, 2018

**Type** Package

**Title** Subclonal copy number and LOH prediction from whole genome sequencing of tumours

**Version** 1.16.0

**Date** 2017-05-10

**Author** Gavin Ha, Sohrab P Shah

**Maintainer** Gavin Ha <gavinha@broadinstitute.org>, Sohrab P Shah <sshah@bccrc.ca>

**Depends** R (>= 3.3.2)

**Imports** IRanges (>= 2.6.1), GenomicRanges (>= 1.24.3), VariantAnnotation (>= 1.18.7), foreach (>= 1.4.3), Rsamtools (>= 1.24.0), GenomeInfoDb (>= 1.8.7), data.table (>= 1.10.4), dplyr (>= 0.5.0),

**Description** Hidden Markov model to segment and predict regions of subclonal copy number alterations (CNA) and loss of heterozygosity (LOH), and estimate cellular prevalence of clonal clusters in tumour whole genome sequencing data.

**License** GPL-3

**biocViews** Sequencing, WholeGenome, DNaseSeq, ExomeSeq, StatisticalMethod, CopyNumberVariation, HiddenMarkovModel, Genetics, GenomicVariation

**URL** <https://github.com/gavinha/TitanCNA>

**NeedsCompilation** yes

## R topics documented:

TitanCNA-package	2
computeSDBwIndex	4
correctReadDepth	6
extractAlleleReadCounts	7
filterData	9
getPositionOverlap	10
haplotype-analysis-methods	12
loadAlleleCounts	14
loadDefaultParameters	16
output-methods	18

Plotting TITAN results . . . . .	21
runEMclonalCN . . . . .	24
TitanCNA trained dataset . . . . .	28
viterbiClonalCN . . . . .	28
WIG Import Functions . . . . .	30

## Index 31

---

TitanCNA-package	<i>TITAN: Subclonal copy number and LOH prediction whole genome sequencing of tumours</i>
------------------	---

---

## Description

TITAN is a software tool for inferring subclonal copy number alterations (CNA) and loss of heterozygosity (LOH). The algorithm also infers clonal group cluster membership for each event and the tumour proportion, or cellular prevalence, for each event.

## Details

Package: TitanCNA  
 Type: Package  
 Version: 1.15.0  
 Date: 2017-05-13  
 License: GPL-3

`example("TitanCNA-package")` for quick tour of functionality and visualization  
`vignette("TitanCNA")` for detailed example

## Author(s)

Gavin Ha, Sohrab P Shah Maintainer: Gavin Ha <gavinha@broadinstitute.org>

## References

Ha, G., Roth, A., Khattra, J., Ho, J., Yap, D., Prentice, L. M., Melnyk, N., McPherson, A., Bashashati, A., Laks, E., Biele, J., Ding, J., Le, A., Rosner, J., Shumansky, K., Marra, M. A., Huntsman, D. G., McAlpine, J. N., Aparicio, S. A. J. R., and Shah, S. P. (2014). TITAN: Inference of copy number architectures in clonal cell populations from tumour whole genome sequence data. *Genome Research*, 24: 1881-1893. (PMID: 25060187)

## Examples

```
message('Running TITAN ...')
#### LOAD DATA ####
infile <- system.file("extdata", "test_alleleCounts_chr2.txt", package = "TitanCNA")
data <- loadAlleleCounts(infile)

#### LOAD PARAMETERS ####
message('titan: Loading default parameters')
numClusters <- 2
```

```

params <- loadDefaultParameters(copyNumber = 5,
                                numberClonalClusters = numClusters, skew = 0.1)

#### READ COPY NUMBER FROM HMMCOPY FILE ####
message('titan: Correcting GC content and mappability biases...')
tumWig <- system.file("extdata", "test_tum_chr2.wig", package = "TitanCNA")
normWig <- system.file("extdata", "test_norm_chr2.wig", package = "TitanCNA")
gc <- system.file("extdata", "gc_chr2.wig", package = "TitanCNA")
map <- system.file("extdata", "map_chr2.wig", package = "TitanCNA")
cnData <- correctReadDepth(tumWig, normWig, gc, map)
logR <- getPositionOverlap(data$chr, data$posn, cnData)
data$logR <- log(2^logR) #transform to natural log

#### FILTER DATA FOR DEPTH, MAPPABILITY, NA, etc ####
data <- filterData(data, c(1:22,"X"), minDepth = 10, maxDepth = 200, map = NULL)

#### EM (FWD-BACK) TO TRAIN PARAMETERS ####
#### Can use parallelization packages ####
K <- length(params$genotypeParams$alphaKHyper)
params$genotypeParams$alphaKHyper <- rep(500, K)
params$ploidyParams$phi_0 <- 1.5
convergeParams <- runEMclonalCN(data, params,
                                maxiter = 3, maxiterUpdate = 500,
                                txnExpLen = 1e9, txnZstrength = 1e9,
                                useOutlierState = FALSE,
                                normalEstimateMethod = "map",
                                estimates = TRUE, estimatePloidy = TRUE)

#### COMPUTE OPTIMAL STATE PATH USING VITERBI ####
optimalPath <- viterbiClonalCN(data, convergeParams)

#### FORMAT RESULTS ####
results <- outputTitanResults(data, convergeParams, optimalPath,
                              filename = NULL, posteriorProbs = FALSE,
                              subcloneProfiles = TRUE,
                              is.haplotypeData = FALSE,
                              correctResults = TRUE,
                              proportionThreshold = 0.05,
                              proportionThresholdClonal = 0.05)
convergeParams <- results$convergeParams
results <- results$corrResults

#### GET SEGMENT RESULTS ####
segs <- outputTitanSegments(results, id = "test", convergeParams,
                             filename = NULL, igvfilename = NULL)

#### PLOT RESULTS ####
norm <- tail(convergeParams$n, 1)
ploidy <- tail(convergeParams$phi, 1)

par(mfrow=c(4, 1))
plotCNlogRByChr(results, chr = 2, segs = segs, ploidy = ploidy, normal = norm, geneAnnot = NULL,
                ylim = c(-2, 2), cex = 0.5, xlab = "", main = "Chr 2")
plotAllelicRatio(results, chr = 2, geneAnnot = NULL, ylim = c(0, 1), cex = 0.5,
                 xlab = "", main = "Chr 2")
plotClonalFrequency(results, chr = 2, normal = norm, geneAnnot = NULL,
                    ylim = c(0, 1), cex = 0.5, xlab = "", main = "Chr 2")
plotSubcloneProfiles(results, chr = 2, cex = 2, main = "Chr 2")

```

```
plotSegmentMedians(segs, chr=2, resultType = "LogRatio", plotType = "CopyNumber",
  plot.new = TRUE, ylim = c(0, 4), main = "Chr 2")
```

---

computeSDBwIndex      *Compute the S\_Dbw Validity Index for **TitanCNA** model selection*

---

## Description

Compute the S\_Dbw Validity Index internal cluster validation from the **TitanCNA** results to use for model selection.

## Usage

```
computeSDBwIndex(x, centroid.method = "median",
  data.type = "LogRatio", S_Dbw.method = "Halkidi",
  symmetric = TRUE)
```

## Arguments

x	Formatted <b>TitanCNA</b> results output from <a href="#">outputTitanResults</a> . See Example.
centroid.method	median or mean method to compute cluster centroids during internal cluster validation.
data.type	Compute S_Dbw validity index based on copy number (use 'LogRatio') or allelic ratio (use 'AllelicRatio').
symmetric	TRUE if the TITAN analysis was carried out using symmetric genotypes. See <a href="#">loadAlleleCounts</a> .
S_Dbw.method	Compute S_Dbw validity index using Halkidi or Tong method. See details and references.

## Details

S\_Dbw Validity Index is an internal clustering evaluation that is used for model selection (Halkidi et al. 2002). It attempts to choose the model that minimizes within cluster variances (scat) and maximizes density-based cluster separation (Dens). Then,  $S\_Dbw(lc\_Tlx\ z) = Dens(lc\_Tlx\ z) + scat(lc\_Tlx\ z)$ .

In the context of **TitanCNA**, if `data.type='LogRatio'`, then the S\_Dbw internal data consists of copy number log ratios, and the resulting joint states of copy number ( $c\_T$ , for all  $c\_T$  in  $\{0 : \max.copy.number\}$ ) and clonal cluster ( $z$ ) make up the clusters in the internal evaluation. If `data.type='AllelicRatio'`, then the S\_Dbw internal data consists of the allelic ratios. The optimal **TitanCNA** run is chosen as the run with the minimum S\_Dbw. If `data.type='Both'`, then the sum of the S\_Dbw for 'LogRatio' and 'AllelicRatio' are added together. This helps account for both data types when choosing the optimal solution.

Note that for S\_Dbw.method, the Tong method has an incorrect formulation of the `scat(c)` function. The function should be a weighted sum, but that is not the formulation shown in the publication. `computeSDBwIndex` uses  $(n_i/N)$  instead of  $(N-n_i)/N$  in the `scat` formula, where  $n_i$  is the number of datapoints in cluster  $i$  and  $N$  is the total number of datapoints.

**Value**

[list](#) with components:

dens.bw	density component of S_Dbw index
scat	scatter component of S_Dbw index
S_DbwIndex	Sum of dens.bw and scat.

**Author(s)**

Gavin Ha <gavinha@gmail.com>

**References**

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2002). Clustering validity checking methods: part ii. SIGMOD Rec., 31(3):19–27.

Tong, J. and Tan, H. Clustering validity based on the improved S\_Dbw\* index. (2009). Journal of Electronics (China), Volume 26, Issue 2, pp 258-264.

Ha, G., Roth, A., Khattra, J., Ho, J., Yap, D., Prentice, L. M., Melnyk, N., McPherson, A., Bashashati, A., Laks, E., Biele, J., Ding, J., Le, A., Rosner, J., Shumansky, K., Marra, M. A., Huntsman, D. G., McAlpine, J. N., Aparicio, S. A. J. R., and Shah, S. P. (2014). TITAN: Inference of copy number architectures in clonal cell populations from tumour whole genome sequence data. Genome Research, 24: 1881-1893. (PMID: 25060187)

**See Also**

[outputModelParameters](#), [loadAlleleCounts](#)

**Examples**

```
data(EMresults)

#### COMPUTE OPTIMAL STATE PATH USING VITERBI ####
#options(cores=1)
optimalPath <- viterbiClonalCN(data, convergeParams)

#### FORMAT RESULTS ####
results <- outputTitanResults(data, convergeParams, optimalPath,
                             filename = NULL, posteriorProbs = FALSE,
                             correctResults = TRUE,
                             proportionThreshold = 0.05,
                             proportionThresholdClonal = 0.05)

results <- results$corrResults ## use corrected results
#### COMPUTE S_Dbw Validity Index FOR MODEL SELECTION ####
s_dbw <- computeSDBwIndex(results, data.type = "LogRatio",
                          centroid.method = "median", S_Dbw.method = "Tong")
```

---

correctReadDepth	<i>Correct GC content and mappability biases in sequencing data read counts</i>
------------------	---

---

### Description

Correct GC content and mappability biases in tumour sequence read counts using Loess curve fitting. Wrapper for function in **HMMcopy**.

### Usage

```
correctReadDepth(tumWig, normWig, gcWig, mapWig,
  genomeStyle = "NCBI", targetedSequence = NULL)
```

### Arguments

tumWig	File path to fixedStep WIG format file for the tumour sample. See <a href="#">wigToRangedData</a> in the <b>HMMcopy</b> for more details.
normWig	File path to fixedStep WIG format file for the normal sample.
gcWig	File path to fixedStep WIG format file for the GC content based on the specific reference genome sequence used.
mapWig	File path to fixedStep WIG format file for the mappability scores computed on the specific reference genome used.
genomeStyle	The genome style to use for chromosomes by <b>TitanCNA</b> . Use one of 'NCBI' or 'UCSC'. It does not matter what style is found in inCounts, genomeStyle will be the style returned.
targetedSequence	data.frame with 3 columns: chr, start position, stop position. Use this argument for exome capture sequencing or targeted deep sequencing data. This is experimental and may not work as desired.

### Details

Wrapper for [correctReadcount](#) in **HMMcopy** package. It uses a sampling of 50000 bins to find the Loess fit. Then, the log ratio for every bin is returned as the log base 2 of the ratio between the corrected tumour read count and the corrected normal read count.

### Value

`data.frame` containing columns:

chr	Chromosome; uses 'X' and 'Y' for sex chromosomes
start	Start genomic coordinate for bin in which read count is corrected
end	End genomic coordinate for bin in which read count is corrected
logR	Log ratio, $\log_2(\text{tumour}:\text{normal})$ , for bin in which read count is corrected

### Author(s)

Gavin Ha <gavinha@gmail.com>, Daniel Lai <jujubix@cs.ubc.ca>, Yikan Wang <ykwang@bccrc.ca>

## References

Ha, G., Roth, A., Lai, D., Bashashati, A., Ding, J., Goya, R., Giuliany, R., Rosner, J., Oloumi, A., Shumansky, K., Chin, S.F., Turashvili, G., Hirst, M., Caldas, C., Marra, M. A., Aparicio, S., and Shah, S. P. (2012). Integrative analysis of genome wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple negative breast cancer. *Genome Research*, 22(10):1995,2007. (PMID: 22637570)

Ha, G., Roth, A., Khattra, J., Ho, J., Yap, D., Prentice, L. M., Melnyk, N., McPherson, A., Bashashati, A., Laks, E., Biele, J., Ding, J., Le, A., Rosner, J., Shumansky, K., Marra, M. A., Huntsman, D. G., McAlpine, J. N., Aparicio, S. A. J. R., and Shah, S. P. (2014). TITAN: Inference of copy number architectures in clonal cell populations from tumour whole genome sequence data. *Genome Research*, 24: 1881-1893. (PMID: 25060187)

## See Also

`correctReadcount` and `wigToRangedData` in the **HMMcopy** package. WIG: <http://genome.ucsc.edu/goldenPath/help/wiggle.html>

## Examples

```
tumWig <- system.file("extdata", "test_tum_chr2.wig", package = "TitanCNA")
normWig <- system.file("extdata", "test_norm_chr2.wig", package = "TitanCNA")
gc <- system.file("extdata", "gc_chr2.wig", package = "TitanCNA")
map <- system.file("extdata", "map_chr2.wig", package = "TitanCNA")

#### GC AND MAPPABILITY CORRECTION ####
cnData <- correctReadDepth(tumWig, normWig, gc, map)
```

---

extractAlleleReadCounts

*Function to extract allele read counts from a sequence alignment (BAM) file*

---

## Description

Function to extract allele read counts from a sequence alignment (BAM) file at specific positions of interest. The positions are passed in as the file path to a file in variant call format (VCF).

## Usage

```
extractAlleleReadCounts(bamFile, bamIndex, positions,
  outputFilename = NULL, pileupParam = PileupParam())
```

## Arguments

bamFile	File path location to the sequencing alignment file (BAM format) from which to extract read counts.
bamIndex	File path location to the BAM index file (usually with extension .bai) corresponding to the sequencing alignment file bamFile.
positions	File path location to the variant call format (VCF) file containing the positions at which read counts are to be extracted.

`outputFilename` If given, will specify the file path to which the result will be output as tab-delimited text. Otherwise, the no output is written to file.

`pileupParam` [PileupParam](#) object from the **Rsamtools**. See Details.

### Details

The `pileupParam` object allows users to specify the sequencing parameters to consider when generating the pileup from which read counts are extracted. This includes `'max_depth'`, `'min_base_quality'`, `'min_mapq'`, `'min_nucleotide_depth'=10` (recommended), `'min_minor_allele_depth'`, `'distinguish_strands'`, `'distinguish_nucleotides'=TRUE` (must be TRUE).

### Value

`data.frame` containing columns:

<code>chr</code>	Chromosome; character
<code>position</code>	Position; numeric
<code>ref</code>	Reference counts; character
<code>refCount</code>	Reference counts; numeric
<code>Nref</code>	Non-reference counts; character
<code>NrefCount</code>	Non-reference counts; numeric

### Author(s)

Gavin Ha <[gavinha@gmail.com](mailto:gavinha@gmail.com)>

### References

Ha, G., Roth, A., Khattra, J., Ho, J., Yap, D., Prentice, L. M., Melnyk, N., McPherson, A., Bashashati, A., Laks, E., Biele, J., Ding, J., Le, A., Rosner, J., Shumansky, K., Marra, M. A., Huntsman, D. G., McAlpine, J. N., Aparicio, S. A. J. R., and Shah, S. P. (2014). TITAN: Inference of copy number architectures in clonal cell populations from tumour whole genome sequence data. *Genome Research*, 24: 1881-1893. (PMID: 25060187)

### See Also

[PileupParam](#); <http://samtools.sourceforge.net/>

### Examples

```
## Not run:
countsDF <- extractAlleleReadCounts(bamFile, bamIndex,
  positions, outputFilename = NULL,
  pileupParam = PileupParam())
data <- loadAlleleCounts(countsDF, symmetric = TRUE,
  genomeStyle = "NCBI")

## End(Not run)
```



---

filterData	<i>Filter list object based on read depth and missing data and returns a filtered <a href="#">data.table</a> object.</i>
------------	--

---

### Description

Filters all vectors in list based on specified chromosome(s) of interest, minimum and maximum read depths, missing data, mappability score threshold

### Usage

```
filterData(data ,chrs = NULL, minDepth = 10, maxDepth = 200,
           positionList = NULL, map = NULL, mapThres = 0.9,
           centromeres = NULL, centromere.flankLength = 0)
```

### Arguments

data	<a href="#">data.table</a> object that contains an arbitrary number of components. Should include 'chr', 'tumDepth'. All vector elements must have the same number of rows where each row corresponds to information pertaining to a chromosomal position.
chrs	character or vector of character specifying the chromosomes to keep. Chromosomes not included in this array will be filtered out. Chromosome style must match the genomeStyle used when running <a href="#">loadAlleleCounts</a>
minDepth	Numeric integer specifying the minimum tumour read depth to include. Positions $\geq$ minDepth are kept.
maxDepth	Numeric integer specifying the maximum tumour read depth to include. Positions $\leq$ maxDepth are kept.
positionList	<a href="#">data.frame</a> with two columns: 'chr' and 'posn'. positionList lists the chromosomal positions to use in the analysis. All positions not overlapping this list will be excluded. Use NULL to use all current positions in data.
map	Numeric array containing map scores corresponding to each position in data. Optional for filtering positions based on mappability scores.
mapThres	Numeric float specifying the mappability score threshold. Only applies if map is specified. map scores $\geq$ mapThres are kept.
centromeres	data.frame containing list of centromere regions. This should contain 3 columns: chr, start, and end. If this argument is used, then data at and flanking the centromeres will be removed.
centromere.flankLength	Integer indicating the length (in base pairs) to the left and to the right of the centromere designated for removal of data.

### Details

All vectors in the input [data.table](#) object, and map, must all have the same number of rows.

### Value

The same [data.table](#) object containing filtered components.

**Author(s)**

Gavin Ha <gavinha@gmail.com>

**References**

Ha, G., Roth, A., Khattra, J., Ho, J., Yap, D., Prentice, L. M., Melnyk, N., McPherson, A., Bashashati, A., Laks, E., Biele, J., Ding, J., Le, A., Rosner, J., Shumansky, K., Marra, M. A., Huntsman, D. G., McAlpine, J. N., Aparicio, S. A. J. R., and Shah, S. P. (2014). TITAN: Inference of copy number architectures in clonal cell populations from tumour whole genome sequence data. *Genome Research*, 24: 1881-1893. (PMID: 25060187)

**See Also**

[loadAlleleCounts](#)

**Examples**

```
infile <- system.file("extdata", "test_alleleCounts_chr2.txt",
                     package = "TitanCNA")
tumWig <- system.file("extdata", "test_tum_chr2.wig", package = "TitanCNA")
normWig <- system.file("extdata", "test_norm_chr2.wig", package = "TitanCNA")
gc <- system.file("extdata", "gc_chr2.wig", package = "TitanCNA")
map <- system.file("extdata", "map_chr2.wig", package = "TitanCNA")

#### LOAD DATA ####
data <- loadAlleleCounts(infile, genomeStyle = "NCBI")

#### GC AND MAPPABILITY CORRECTION ####
cnData <- correctReadDepth(tumWig, normWig, gc, map)

#### READ COPY NUMBER FROM HMMCOPY FILE ####
logR <- getPositionOverlap(data$chr, data$posn, cnData)
data$logR <- log(2*logR) #use natural logs

#### FILTER DATA FOR DEPTH, MAPPABILITY, NA, etc ####
filtereData <- filterData(data, as.character(1:24), minDepth = 10,
                          maxDepth = 200, map = NULL, mapThres=0.9)
```

---

getPositionOverlap	<i>Function to assign values to given chromosome-position that overlaps a list of chromosomal segments</i>
--------------------	--

---

**Description**

Given a list of chromosomes and positions, uses a C-based function that searches a list of segments to find the overlapping segment. Then, takes the value (4th column in segment data.frame) of the overlapping segment and assigns to the given chromosome and position.

**Usage**

```
getPositionOverlap(chr, posn, dataVal)
```

**Arguments**

chr	Numeric <a href="#">array</a> denoting the chromosome for a list of positions. Must have the same number of elements as posn.
posn	Numeric <a href="#">array</a> denoting the position in the chromosome for a list of positions. Must have the same number of elements as chr.
dataVal	<a href="#">data.frame</a> containing a list of segments described with 4 columns: chromosome, start coordinate, end coordinate, value of interest (e.g. log ratio). Chromosome can be all numeric or chrX and chrY can use 'X' and 'Y'.

**Value**

Numeric [array](#) of values from the 4th column of [data.frame](#) cnData. Each element corresponds to a genomic location from chr and posn that overlapped the segment in cnData.

**Author(s)**

Gavin Ha <gavinha@gmail.com>

**References**

Ha, G., Roth, A., Khattra, J., Ho, J., Yap, D., Prentice, L. M., Melnyk, N., McPherson, A., Bashashati, A., Laks, E., Biele, J., Ding, J., Le, A., Rosner, J., Shumansky, K., Marra, M. A., Huntsman, D. G., McAlpine, J. N., Aparicio, S. A. J. R., and Shah, S. P. (2014). TITAN: Inference of copy number architectures in clonal cell populations from tumour whole genome sequence data. *Genome Research*, 24: 1881-1893. (PMID: 25060187)

**See Also**

[loadAlleleCounts](#), [correctReadDepth](#)

**Examples**

```
infile <- system.file("extdata", "test_alleleCounts_chr2.txt",
                     package = "TitanCNA")
tumWig <- system.file("extdata", "test_tum_chr2.wig", package = "TitanCNA")
normWig <- system.file("extdata", "test_norm_chr2.wig", package = "TitanCNA")
gc <- system.file("extdata", "gc_chr2.wig", package = "TitanCNA")
map <- system.file("extdata", "map_chr2.wig", package = "TitanCNA")

#### LOAD DATA ####
data <- loadAlleleCounts(infile)

#### GC AND MAPPABILITY CORRECTION ####
cnData <- correctReadDepth(tumWig, normWig, gc, map)

#### READ COPY NUMBER FROM HMMCOPY FILE ####
logR <- getPositionOverlap(data$chr, data$posn, cnData)
```

---

 haplotype-analysis-methods

*Function to load tumour allele counts from a text file or data.frame and returns a [data.table](#) (loadHaplotypeAlleleCounts).  
Function to load phased heterozygous sites from a VCF file (getHaplotypesFromVCF)*

---

**Description**

Function to load in the allele counts from tumour sequencing data from a delimited text file or data.frame object.

**Usage**

```
loadHaplotypeAlleleCounts(inCounts, fun = "sum", haplotypeBinSize = 1e5,
  minSNPsInBin = 3, chrs = c(1:22, "X"), minNormQual = 200,
  genomeStyle = "NCBI", sep = "\t", header = TRUE, seqinfo = NULL)

getHaplotypesFromVCF(vcfFile, chrs = c(1:22, "X"), build = "hg19",
  filterFlags = c("PASS", "10X_RESCUED_MOLECULE_HIGH_DIVERSITY"),
  minQUAL = 100, minDepth = 10, minVAF = 0.25, altCountField = "AD",
  keepGenotypes = c("1|0", "0|1", "0/1"), snpDB = NULL)
```

**Arguments**

- |                  |  |
|------------------|--|
| inCounts         | Full file path to text file or data.frame containing tumour allele count data. inCounts must be 6 columns: chromosome, position, reference base, reference read counts, non-reference base, non-reference read counts. 'chromosome' column can be in 'NCBI' or 'UCSC' genome style; only autosomes, sex chromosomes, and mitochondrial chromosome are included (e.g. 1-22,X,Y,MT). The reference and non-reference base columns can be any arbitrary character; it is not used by <b>TitanCNA</b> .  |
| vcfFile          | Path to phased variant VCF file from LongRanger 2.1. The file name must have the suffix *phase_variants.vcf.gz.  |
| fun              | The function ('SNP', 'sum', 'mean') to use to summarize within each user defined bin using haplotypeBinSize and haplotype block defined by the phaseSet ID from the 9th column of inCounts. 'SNP' - uses the phased allele counts each individual SNP; phased allele for the higher coverage (determined within each bin) haplotype is chosen. 'sum' - uses the read count sum across all phased SNPs for the higher coverage haplotype within a bin normalized by the total depth across all SNPs in a bin; each SNP in the bin is assigned this fraction. 'mean' - uses the mean (rounded) read count across all phased SNPs for the higher coverage haplotype within a bin normalized by the mean (rounded) depth across all SNPs in a bin; each SNP in the bin is assigned this rounded count and depth. |
| haplotypeBinSize | Bin size used to summarize SNPs based on phased haplotypes. See fun for the summarization approaches within a bin.   |
| minSNPsInBin     | The minimum number of SNPs required in each haplotypeBinSize for analysis. See fun for the summarization approaches within a bin.  |

chrs	Vector containing list of chromosomes to include in output.
minNormQual	Quality threshold to use for filtering; SNPs with lower than this value are excluded. This quality is any metric that provides the confidence of the locus being a true germline heterozygous SNP.
genomeStyle	The genome style to use for chromosomes. Use one of 'NCBI' or 'UCSC'. It does not matter what style is found in inCounts, genomeStyle will be the style returned. Invokes <a href="#">setGenomeStyle</a> .
build	Human genome reference build. Default: hg19.
snpDB	Path to SNP VCF file to use for specifying sites to retain.
minQUAL	Variants with quality (QUAL field) greater or equal to this value will be retained.
minDepth	Variants with read depth greater than or equal to this value will be retained.
minVAF	Variants with a variant/reference allele fraction of greater than or equal to this value will be retained.
altCountField	Specify the alternate count field name. Default is "AD".
keepGenotypes	Genotypes to retain. Default is to keep these genotypes strings: 1/0, 0/1, 0/1
filterFlags	Specify the FILTER flags to retain.
sep	Character indicating the delimiter used for the columns for infile. Default is tab-delimited, "\t".
header	logical to indicate if the input tumour counts file contains a header line.
seqinfo	<a href="#">Seqinfo-class</a> object describing chromosome information. If NULL, then will load seqinfo for hg19 system.files('extdata', 'Seqinfo_hg19.rda', package='TitanCNA').

### Value

loadHaplotypeAlleleCounts returns a [data.table](#) containing components for

chr	Chromosome; character, genomeStyle naming convention
posn	Position; integer
phaseSet	Phase block identifier, numeric or character
refOriginal	Reference allele read count at SNP; numeric
tumDepthOriginal	Coverage at SNP; numeric
ref	Phased allele count values of higher coverage haplotype based on approach used (SNP, sum, mean); numeric
nonRef	Phased allele count values of lower coverage haplotype; tumDepth minus ref; numeric
tumDepth	Mean or sum of SNP read coverage; numeric
HaplotypeRatio	Sum of read coverage of phased alleles of higher coverage haplotype normalized by tumDepth; numeric
haplotypeCount	Phased allele read count; numeric

getHaplotypesFromVCF returns a [list](#) containing 2 components

vcf.filtered	VCF object containing the list of heterozygous variants after filtering.
geno.gr	GRanges object containing the genotype information of the VCF

**Author(s)**

Gavin Ha <gavinha@gmail.com>

**References**

Ha, G., Roth, A., Khattra, J., Ho, J., Yap, D., Prentice, L. M., Melnyk, N., McPherson, A., Bashashati, A., Laks, E., Biele, J., Ding, J., Le, A., Rosner, J., Shumansky, K., Marra, M. A., Huntsman, D. G., McAlpine, J. N., Aparicio, S. A. J. R., and Shah, S. P. (2014). TITAN: Inference of copy number architectures in clonal cell populations from tumour whole genome sequence data. *Genome Research*, 24: 1881-1893. (PMID: 25060187)

**See Also**

[loadDefaultParameters](#), [plotHaplotypeFraction](#)

**Examples**

```
## Not run:
infile <- "test_alleleCounts_chr2_with_phaseInfo.txt"
haplotypeBinSize <- 1e5
phaseSummarizeFun <- "sum"
## will load seqinfo_hg19 provided by TitanCNA package
data <- loadHaplotypeAlleleCounts(infile, fun = phaseSummarizeFun,
  haplotypeBinSize = haplotypeBinSize, minSNPsInBin = 3,
  chrs = c(1:22, "X"), minNormQual = 200,
  genomeStyle = "NCBI", seqinfo = NULL)

## End(Not run)

## Not run:
vcfFile <- "test.vcf"
hap <- getHaplotypesFromVCF(vcfFile, chrs = c(1:22,"X"), build = "hg19",
  filterFlags = c("PASS", "10X_RESCUED_MOLECULE_HIGH_DIVERSITY"),
  minQUAL = 100, minDepth = 10, minVAF = 0.25,
  keepGenotypes = ("1|0", "0|1", "0/1"))

## End(Not run)
```

---

loadAlleleCounts	<i>Function to load tumour allele counts from a text file or data.frame and returns a <a href="#">data.table</a>.</i>
------------------	---

---

**Description**

Function to load in the allele counts from tumour sequencing data from a delimited text file or data.frame object.

**Usage**

```
loadAlleleCounts(inCounts, symmetric = TRUE,
  genomeStyle = "NCBI", sep = "\t", header = TRUE)

setGenomeStyle(x, genomeStyle = "NCBI", species = "Homo_sapiens")
```

**Arguments**

inCounts	Full file path to text file or data.frame containing tumour allele count data. inCounts must be 6 columns: chromosome, position, reference base, reference read counts, non-reference base, non-reference read counts. 'chromosome' column can be in 'NCBI' or 'UCSC' genome style; only autosomes, sex chromosomes, and mitochondrial chromosome are included (e.g. 1-22,X,Y,MT). The reference and non-reference base columns can be any arbitrary character; it is not used by <b>TitanCNA</b> .
symmetric	logical; if TRUE, then the symmetric allelic counts will be used. ref will equal max(ref, nonRef). This parameter must be the same as the symmetric parameter for <a href="#">loadDefaultParameters</a> .
genomeStyle	The genome style to use for chromosomes by <b>TitanCNA</b> . Use one of 'NCBI' or 'UCSC'. It does not matter what style is found in inCounts, genomeStyle will be the style returned.
sep	Character indicating the delimiter used for the columns for infile. Default is tab-delimited, "\t".
header	logical to indicate if the input tumour counts file contains a header line.
x	character vector of chromosome names to change.
species	character denoting the species

**Value**

loadAlleleCounts returns a [data.table](#) containing components for

chr	Chromosome; character, NCBI or UCSC genome style format
posn	Position; integer
ref	Reference counts; numeric
nonRef	Non-reference counts; numeric
tumDepth	Tumour depth; numeric

**Author(s)**

Gavin Ha <gavinha@gmail.com>

**References**

Ha, G., Roth, A., Khattra, J., Ho, J., Yap, D., Prentice, L. M., Melnyk, N., McPherson, A., Bashashati, A., Laks, E., Biele, J., Ding, J., Le, A., Rosner, J., Shumansky, K., Marra, M. A., Huntsman, D. G., McAlpine, J. N., Aparicio, S. A. J. R., and Shah, S. P. (2014). TITAN: Inference of copy number architectures in clonal cell populations from tumour whole genome sequence data. *Genome Research*, 24: 1881-1893. (PMID: 25060187)

**See Also**

[loadDefaultParameters](#)

**Examples**

```

infile <- system.file("extdata", "test_alleleCounts_chr2.txt",
                      package = "TitanCNA")
#### LOAD DATA FROM TEXT FILE ####
data <- loadAlleleCounts(infile, symmetric = TRUE,
                        genomeStyle = "NCBI", header = TRUE)

## use the UCSC chromosome naming convention instead ##
data$chr <- setGenomeStyle(data$chr, genomeStyle = "UCSC")
## Not run:
data <- loadAlleleCounts(countsDF, symmetric = TRUE,
                        genomeStyle = "NCBI")

## End(Not run)

```

---

loadDefaultParameters *Load TITAN parameters*

---

**Description**

Load TITAN model parameters based on maximum copy number and number of clonal clusters.

**Usage**

```

loadDefaultParameters(copyNumber = 5, numberClonalClusters = 1, skew = 0,
                      hetBaselineSkew = NULL, alleleEmissionModel = "binomial",
                      symmetric = TRUE, data = NULL)

```

**Arguments**

copyNumber	Maximum number of absolute copies to account for in the model. Default (and recommended) is 5.
numberClonalClusters	Number of clonal clusters to use in the analysis. Each cluster represents a potential clone. Using '1' treats the sample as being clonal (no subclonality). '2' or higher treats the tumour data as being subclonal.
skew	numeric float (between 0 to 0.5) indicating the heterozygous baseline shift for the allelic ratios towards 1. This is may be required for SOLiD data, but for most cases, this argument can be omitted. Use 0 or NULL for no skew.
hetBaselineSkew	Allelic reference base skew for heterozygous states (e.g. 1:1, 2:2, 3:3). Value is additive to baseline allelic ratios (which may already be adjusted by skew). Use 0 or NULL for no skew; use from range between 0 to 0.5.
alleleEmissionModel	Specify the emission model to use for the allelic input data. "binomial" or "Gaussian".
symmetric	logical; if TRUE, then treat genotypes as symmetric. This should always be TRUE because symmetric=FALSE is deprecated. See Details.
data	data is the output of function <a href="#">loadAlleleCounts</a> . If provided and symmetric=TRUE, then it will compute the median allelic ratio to use as the baseline for heterozygous genotypes; otherwise, the baseline will default to 0.55 (reference/depth) if data=NULL. If symmetric=FALSE, this argument will not be used.





---

 output-methods

*Formatting and printing **TitanCNA** results.*


---

### Description

Function to format **TitanCNA** results in to a data.frame and output the results to a tab-delimited file.

### Usage

```
outputTitanResults(data, convergeParams, optimalPath, filename = NULL,
  is.haplotypeData = FALSE, posteriorProbs = FALSE, subcloneProfiles = TRUE,
  correctResults = TRUE, proportionThreshold = 0.05,
  proportionThresholdClonal = 0.05, recomputeLogLik = TRUE, rerunViterbi = FALSE,
  verbose = TRUE)
```

```
outputModelParameters(convergeParams, results, filename,
  S_Dbw.scale = 1, S_Dbw.method = "Tong")
```

```
outputTitanSegments(results, id, convergeParams, filename = NULL,
  igvfilename = NULL)
```

### Arguments

- |                  |   |
|------------------|---|
| id               | Character string identifier for sample  |
| data             | <a href="#">list</a> object that contains the components for the data to be analyzed. chr, posn, ref, and tumDepth that can be obtained using <a href="#">loadAlleleCounts</a> , and logR that can be obtained using <a href="#">correctReadDepth</a> and <a href="#">getPositionOverlap</a> (see Example). |
| convergeParams   | <a href="#">list</a> object that is returned from the function <a href="#">runEMclonalCN</a> in <b>TitanCNA</b> .   |
| optimalPath      | numeric <a href="#">array</a> containing the optimal <b>TitanCNA</b> genotype and clonal cluster states for each data point in the analysis. optimalPath is obtained from running <a href="#">viterbiClonalCN</a> .   |
| results          | Formatted <b>TitanCNA</b> results output from <a href="#">outputTitanResults</a> .  |
| filename         | Path of the file to write the <b>TitanCNA</b> results.  |
| igvfilename      | Path of the file to write the IGV seg file.   |
| posteriorProbs   | Logical TRUE to include the posterior marginal probabilities in printing to filename.   |
| is.haplotypeData | Logical TRUE if the data contains the haplotype information. In particular, the column headers HaplotypeCount, HaplotypeDepth, HaplotypeRatio are included.   |
| subcloneProfiles | Logical TRUE to include the subclone profiles to the output data.frame. Currently, this only works for 1 or 2 clonal clusters.  |
| correctResults   | Logical TRUE to correct the results by removing empty clusters and adjusting cellular prevalence and normal contamination parameters accordingly.   |

recomputeLogLik	Logical TRUE to re-run forwards-backwards to re-estimate the log-likelihood after correcting results (e.g. correctResults is TRUE)
rerunViterbi	Logical TRUE to re-run viterbi to segment the results again after correcting results (e.g. correctResults is TRUE)
proportionThreshold	Minimum proportion of the genome altered (by SNPs) for a cluster to be retained. Clonal clusters having lower proportion of alteration are removed.
proportionThresholdClonal	Minimum proportion of genome altered by clonal events (by SNPs) for the highest cellular prevalence cluster. If the highest prevalence cluster contains lower proportion of events than this threshold, this cluster will be removed and the next highest (subclonal) cluster will be readjusted to be the clonal cluster.
S_Dbw.scale	The S_Dbw validity index can be adjusted to account for differences between datasets. S_Dbw.scale can be used to penalize the S_Dbw dens.bw component. The default is 1.
S_Dbw.method	Compute S_Dbw validity index using Halkidi or Tong method. See <a href="#">computeS_DbwIndex</a> .
verbose	Print status messages.

## Details

[outputModelParameters](#) outputs to a file with the estimated TITAN model parameters and model selection index. Each row contains information regarding different parameters:

- 1) Normal contamination estimate - proportion of normal content in the sample; tumour content is 1 minus this number
- 2) Average tumour ploidy estimate - average number of estimated copies in the genome; 2 represents diploid
- 3) Clonal cluster cellular prevalence - Z denotes the number of clonal clusters; each value (space-delimited) following are the cellular prevalence estimates for each cluster. Cellular prevalence here is defined as the proportion of tumour sample that does contain the aberrant genotype.
- 4) Genotype binomial means for clonal cluster Z - set of 21 binomial estimated parameters for each specified cluster
- 5) Genotype Gaussian means for clonal cluster Z - set of 21 Gaussian estimated means for each specified cluster
- 6) Genotype Gaussian variance - set of 21 Gaussian estimated variances; variances are shared for across all clusters
- 7) Number of iterations - number of EM iterations needed for convergence
- 8) Log likelihood - complete data log-likelihood for current cluster run
- 9) S\_Dbw dens.bw - density component of S\_Dbw index; see [computeS\\_DbwIndex](#)
- 10) S\_Dbw scat - scatter component of S\_Dbw index; see [computeS\\_DbwIndex](#)
- 11) S\_Dbw validity index - used for model selection where the run with optimal number of clusters based on lowest S\_Dbw index. This value is slightly modified from that computed from [computeS\\_DbwIndex](#). It is computed as  $S\_Dbw = S\_Dbw.scale * dens.bw + scat$
- 12) S\_Dbw dens.bw, scat, validity index is computed for LogRatio and AllelicRatio datatypes, as well as the combination of Both. For Both, the values are summed for both datatypes.

[outputTitanResults](#) outputs a file that has the similar format described in 'Value' section.

**Value**

`outputTitanResults` also returns a list containing the following:

<code>results</code>	TITAN results, uncorrected for cluster number and parameters
<code>corrResults</code>	TITAN results, corrected by removing empty clusters and parameters adjusted accordingly.
<code>convergeParams</code>	Corrected parameter object

The `results` and `corrResults` are `data.table` objects, where each row corresponds to a position in the analysis, and with the following columns:

<code>Chr</code>	character denoting chromosome number. ChrX and ChrY uses 'X' and 'Y'.
<code>Position</code>	genomic coordinate
<code>RefCount</code>	number of reads matching the reference base
<code>NRefCount</code>	number of reads matching the non-reference base
<code>Depth</code>	total read depth at the position
<code>AllelicRatio</code>	<code>RefCount/Depth</code>
<code>LogRatio</code>	<code>log2</code> ratio between normalized tumour and normal read depths
<code>CopyNumber</code>	predicted <b>TitanCNA</b> copy number
<code>TITANstate</code>	internal state number used by <b>TitanCNA</b> ; see Reference
<code>TITANcall</code>	interpretable <b>TitanCNA</b> state; string (HOMD,DLOH,HET,NLOH,ALOH,ASCNA,BCNA,UBCNA); See Reference
<code>ClonalCluster</code>	predicted <b>TitanCNA</b> clonal cluster; lower cluster numbers represent clusters with higher cellular prevalence
<code>CellularPrevalence</code>	proportion of tumour cells containing event; not to be mistaken as proportion of sample (including normal)

If `subcloneProfiles` is set to `TRUE`, then the subclone profiles will be appended to the output `data.frame`.

<code>Subclone1.CopyNumber</code>	Integer copy number for Subclone 1.
<code>Subclone1.TITANcall</code>	States for Subclone 1
<code>Subclone1.Prevalence</code>	The cellular prevalence of Subclone 1, or sometimes referred to as the subclone fraction.

`outputModelParameters` returns a `list` containing the `S_Dbw` model selection:

<code>dens.bw</code>	
<code>scat</code>	
<code>S_Dbw</code>	<code>S_Dbw.scale * dens.bw + scat</code>

**Author(s)**

Gavin Ha <gavinha@gmail.com>

## References

Ha, G., Roth, A., Khattra, J., Ho, J., Yap, D., Prentice, L. M., Melnyk, N., McPherson, A., Bashashati, A., Laks, E., Biele, J., Ding, J., Le, A., Rosner, J., Shumansky, K., Marra, M. A., Huntsman, D. G., McAlpine, J. N., Aparicio, S. A. J. R., and Shah, S. P. (2014). TITAN: Inference of copy number architectures in clonal cell populations from tumour whole genome sequence data. *Genome Research*, 24: 1881-1893. (PMID: 25060187)

## See Also

[runEMclonalCN](#), [viterbiClonalCN](#), [computeSDBwIndex](#)

## Examples

```
data(EMresults)

#### COMPUTE OPTIMAL STATE PATH USING VITERBI ####
optimalPath <- viterbiClonalCN(data, convergeParams)

#### FORMAT RESULTS ####
results <- outputTitanResults(data, convergeParams, optimalPath,
                              filename = NULL, posteriorProbs = FALSE,
                              subcloneProfiles = TRUE, correctResults = TRUE,
                              proportionThreshold = 0.05, recomputeLogLik = FALSE,
                              proportionThresholdClonal = 0.05,
                              is.haplotypeData = FALSE)

## use corrected parameters
convergeParams <- results$convergeParam
## use corrected results
results <- results$corrResults

#### OUTPUT RESULTS TO FILE ####
outparam <- paste0("cluster2_params.txt")
outputModelParameters(convergeParams, results, outparam)

#### OUTPUT SEGMENTS TO FILE ####
outseg <- paste0("cluster2_segs.txt")
outigv <- paste0("cluster2_seg")
segs <- outputTitanSegments(results, id = "test", convergeParams,
                             filename = outseg, igvfilename = outigv)
# segment results also stored in data.frame "segs"
```

---

Plotting TITAN results

*Plotting functions for **TitanCNA** results.*

---

## Description

Three plotting functions for **TitanCNA** results. `plotCNlogRByChr` plots the copy number results from log ratio data. `plotAllelicRatio` plots the allelic imbalance and loss of heterozygosity (LOH) from allelic ratio data. `plotClonalFrequency` plots the clonal cluster and cellular prevalence results for each data point.

**Usage**

```

plotAllelicRatio(dataIn, chr = NULL, geneAnnot = NULL, spacing = 4,
  xlim = NULL, ...)
plotClonalFrequency(dataIn, chr = NULL, normal = NULL, geneAnnot = NULL,
  spacing = 4, xlim = NULL, ...)
plotCNlogRByChr(dataIn, chr = NULL, segs = NULL, geneAnnot = NULL, ploidy = NULL,
  normal = NULL, spacing = 4, alphaVal = 1, xlim = NULL, ...)
plotSubcloneProfiles(dataIn, chr = NULL, geneAnnot = NULL,
  spacing = 4, xlim = NULL, ...)
plotSegmentMedians(dataIn, resultType = "LogRatio", plotType = "CopyNumber",
  chr = NULL, geneAnnot = NULL, ploidy = NULL, spacing = 4, alphaVal = 1, xlim = NULL,
  plot.new = FALSE, lwd = 8, ...)
plotHaplotypeFraction(dataIn, chr = NULL, resultType = "HaplotypeRatio", colType = "Haplotypes",
  geneAnnot = NULL, spacing = 4, xlim = NULL, ...)

```

**Arguments**

dataIn	Formatted <b>TitanCNA</b> results output from <a href="#">outputTitanResults</a> . See Example.
chr	Plot results for specified chr. If chr is NULL, then results for the entire genome is plot.
segs	<a href="#">data.frame</a> containing named columns: Chromosome, Median_logR, Start_Position.bp., End_Position.bp.. This data can be read in from the segments generated by the TITANRunner pipeline. These segments will be overlaid in the plots as lines at the median log ratio for each segment.
resultType	For plotSegmentMedians: specify the data type ('LogRatio' or 'AllelicRatio') to plot. For plotHaplotypeFraction: specify the data type ('HaplotypeRatio' or 'AllelicRatio') to plot.
plotType	Specify whether to plot the 'CopyNumber' or 'Ratio' values for the resultType.
colType	Specify the color scheme to use: 'Haplotypes' or 'CopyNumber'. For 'Haplotypes', alternating blue and red used to illustrated the data within phased haplotype blocks. For 'CopyNumber', the same colors as plotAllelicRatio are used for allelic copy number events.
geneAnnot	<a href="#">data.frame</a> specifying the genes to annotate in the plot. Gene boundaries are indicated using vertical dotted grey lines and gene symbols are shown at the top of the plot. geneAnnot must have four columns: gene symbol, chr, start coordinate, stop coordinate.
normal	numeric scalar indicating the normal contamination. This can be obtained from converge parameters output using <a href="#">runEMclonalCN</a> . See Example.
ploidy	numeric scalar indicating the tumour ploidy used to adjust the copy number plot <a href="#">plotCNlogRByChr</a> . This can be obtained from converge parameters output using <a href="#">runEMclonalCN</a> . See Example. If NULL is used, then ploidy adjustment is not used in the plot.
spacing	Number of lines of spacing for the margin spacing at the bottom of the plot. Useful if an idiogram/karogram is plot underneath.
alphaVal	Set an alpha value between 0 and 1 to allow transparency in the points being plot.
xlim	Two element vector to specify the xlim for the plot. If NULL, then entire chromosome is plot.
lwd	Explicitly specify the line width for segments being plot.

`plot.new` TRUE if a new plot is used. Set to FALSE to overlay an existing plot.  
 ... Additional arguments used in the `plot` function.

## Details

`plotCNlogRByChr` plots the copy number alterations from log ratio data. The Y-axis is based on log ratios. Log ratios are computed ratios between normalized tumour and normal read depths. Data points close to 0 represent diploid, above 0 are copy gains, below 0 are deletions. `ploidy` argument adjusts the baseline of the data points. Colours represent the copy number state. Bright Green - Homozygous deletion (HOMD) Green - Hemizygous deletion (DLOH) Blue - Diploid heterozygous (HET), Copy-neutral LOH (NLOH) Dark Red - GAIN Red - Allele-specific CNA (ASCNA), Unbalanced CNA (UBCNA), Balanced CNA (BCNA)

`plotAllelicRatio` plots the allelic imbalance and loss of heterozygosity from allelic ratio data. The Y-axis is based on allelic ratios. Allelic ratios are computed as  $\text{RefCount/Depth}$ . Data points close to 1 represent homozygous reference base, close to 0 represent homozygous non-reference base, and close to 0.5 represent heterozygous. Normal contamination influences the divergence away from 0.5 for LOH events. No adjustments are made to the plot as the original data from `dataIn` are shown. Colours represent the allelic imbalance and LOH state. Grey - HET, BCNA Bright Green - HOMD Green - DLOH, ALOH Blue - NLOH Dark Red - GAIN Red - ASCNA, UBCNA

`plotClonalFrequency` plots the cellular prevalence and clonal clusters from the results. The Y-axis is the cellular prevalence that includes the normal proportion. Therefore, the cellular prevalence here refers to the proportion in the sample (including normal). Lines are drawn for each data point indicating the cellular prevalence. Heterozygous diploid are not shown because it is a normal genotype and is not categorized as being subclonal (this means 100% of cells are normal). The black horizontal line represents the tumour content labeled as 'T'. Each horizontal grey line represents the cellular prevalence of the clonal clusters labeled as Z1, Z2, etc. Colours are the same for allelic ratio plots.

`plotSubcloneProfiles` plots the predicted copy number profile for individual subclones inferred by TITAN. Currently, this only works for solutions having 1 or 2 clonal clusters. The colours are the same as for `plotAllelicRatio`.

`plotSegmentMedians` plots the segment means for 'LogRatio' or 'AllelicRatio' data. There are also two types of plots for each of the datatypes: 'Ratio' or 'CopyNumber'. For 'Ratio', the logRatio or allelic ratios in the output results files are plot. For 'CopyNumber', the y-axis is converted to the exponentiated absolute copy number levels for the easier interpretability of the results.

`plotHaplotypeFraction` plots the phased SNP read count of the higher coverage haplotype, normalized by the total coverage of the haplotype. For 'Haplotypes', alternating colors of blue and red are used to illustrate the phased haplotype blocks provided from the input data (see `loadHaplotypeAlleleCounts`).

## Author(s)

Gavin Ha <gavinha@gmail.com>

## References

Ha, G., Roth, A., Khattra, J., Ho, J., Yap, D., Prentice, L. M., Melnyk, N., McPherson, A., Bashashati, A., Laks, E., Biele, J., Ding, J., Le, A., Rosner, J., Shumansky, K., Marra, M. A., Huntsman, D. G., McAlpine, J. N., Aparicio, S. A. J. R., and Shah, S. P. (2014). TITAN: Inference of copy number architectures in clonal cell populations from tumour whole genome sequence data. *Genome Research*, 24: 1881-1893. (PMID: 25060187)

**See Also**

[outputTitanResults](#), [runEMclonalCN](#), [computeSdbwIndex](#)

**Examples**

```
data(EMresults)

#### COMPUTE OPTIMAL STATE PATH USING VITERBI ####
optimalPath <- viterbiClonalCN(data, convergeParams)

#### FORMAT RESULTS ####
results <- outputTitanResults(data, convergeParams, optimalPath,
  filename = NULL, posteriorProbs = FALSE,
  correctResults = TRUE, proportionThreshold = 0.05,
  proportionThresholdClonal = 0.05)
convergeParams <- results$convergeParams
results <- results$corrResults # use corrected results
#### PLOT RESULTS ####
norm <- tail(convergeParams$n, 1)
ploidy <- tail(convergeParams$phi, 1)

par(mfrow=c(4, 1))
plotCNlogRByChr(results, chr = 2, segs = NULL, ploidy = ploidy, normal = norm,
  geneAnnot = NULL, ylim = c(-2, 2), cex = 0.5, xlab = "",
  main = "Chr 2")
plotAllelicRatio(results, chr = 2, geneAnnot = NULL, ylim = c(0, 1), cex = 0.5,
  xlab = "", main = "Chr 2")
plotClonalFrequency(results, chr = 2, normal = norm, geneAnnot = NULL,
  ylim = c(0, 1), cex = 0.5, xlab = "", main = "Chr 2")
plotSubcloneProfiles(results, chr = 2, cex = 2, main = "Chr 2")

segs <- outputTitanSegments(results, id = "test", convergeParams)

plotSegmentMedians(segs, chr=2, resultType = "LogRatio",
  plotType = "CopyNumber", plot.new = TRUE)
```

---

runEMclonalCN

*Function to run the Expectation Maximization Algorithm in TitanCNA.*

---

**Description**

Function to run the Expectation Maximization Algorithm for inference of model parameters: cellular prevalence, normal proportion, tumour ploidy. This is a key function in the **TitanCNA** package and is the most computationally intense. This function makes calls to a C subroutine that allows the algorithm to be run more efficiently.

**Usage**

```
runEMclonalCN(data, params,
  txnExpLen = 1e15, txnZstrength = 5e05, maxiter = 15,
  maxiterUpdate = 1500, pseudoCounts = 1e-300,
```



```
normalEstimateMethod = "map", estimateS = TRUE,
estimatePloidy = TRUE, useOutlierState = FALSE,
likChangeThreshold = 0.001, verbose = TRUE)
```

## Arguments

data	<b>list</b> object that contains the components for the data to be analyzed. chr, posn, ref, and tumDepth that can be obtained using <code>loadAlleleCounts</code> , and logR that can be obtained using <code>correctReadDepth</code> and <code>getPositionOverlap</code> (see Example).
params	<b>list</b> object that contains major parameters: <b>list</b> object containing copy number and allelic ratio genotype parameters. <b>list</b> object containing the normal contamination parameters. <b>list</b> object containing the tumour ploidy parameters. <b>list</b> object containing the subclonality (cellular prevalence and clonal cluster) parameters. params can be obtained from <code>loadDefaultParameters</code> .
txnExpLen	Influences prior probability of genotype transitions in the HMM. Smaller value have lower tendency to change state; however, too small and it produces underflow problems. 1e-9 works well for up to 3 million total positions.
txnZstrength	Influences prior probability of clonal cluster transitions in the HMM. Smaller value have lower tendency to change clonal cluster state. 1e-9 works well for up to 3 million total positions.
pseudoCounts	Small, machine precision values to add to probabilities to avoid underflow. For example, <code>.Machine\$double.eps</code> .
maxiter	Maximum number of expectation-maximization iterations allowed. In practice, for <b>TitanCNA</b> , it will usually not exceed 20.
maxiterUpdate	Maximum number of coordinate descent iterations during the M-step (of EM algorithm) when parameters are estimated.
normalEstimateMethod	Specifies how to handle normal proportion estimation. Using <code>map</code> will use the maximum a posteriori estimation. Using <code>fixed</code> will not estimate the normal proportion; the normal proportion will be fixed to whatever is specified in <code>params\$normalParams\$n_0</code> . See Details.
estimateS	Logical indicating whether to account for clonality and estimate subclonal events. See Details.
estimatePloidy	Logical indicating whether to estimate and account for tumour ploidy.
useOutlierState	Logical indicating whether an additional outlier state should be used. In practice, this is usually not necessary.
likChangeThreshold	EM convergence criteria - stop EM when complete log likelihood changes less than the proportion specified by this argument.
verbose	Set to <code>FALSE</code> to suppress program messages.

## Details

This function is implemented with the "`foreach`" package and therefore supports parallelization. See "`doMC`" or "`doMPI`" for some parallelization packages.

The forwards-backwards algorithm is used for the E-step in the EM algorithm. This is done using a call to a C subroutine for each chromosome. The maximization step uses maximum a posteriori (MAP) for estimation of parameters.

If the sample has absolutely no normal contamination, then assign `nParams$n_0 <- 0` and use argument `normalEstimateMethod="fixed"`.

`estimateS` should always be set to `TRUE`. If no subclonality is expected, then use `loadDefaultParameters(numberClonalClusters=0)`. Using `estimateS=FALSE` and `loadDefaultParameters(numberClonalClusters=0)` gives more or less the same results.

## Value

`list` with components for results returned from the EM algorithm, including converged parameters, posterior marginal responsibilities, log likelihood, and original parameter settings.

<code>n</code>	Converged estimate for normal contamination parameter. numeric array containing estimates at each EM iteration.
<code>s</code>	Converged estimate(s) for cellular prevalence parameter(s). This value is defined as the proportion of tumour sample that does <i>not</i> contain the aberrant genotype. This will contrast what is output in <code>outputTitanResults</code> . numeric array containing estimates at each EM iteration. If more than one cluster is specified, then <code>s</code> is a numeric matrix.
<code>var</code>	Converged estimates for variance parameter of the Gaussian mixtures used to model the log ratio data. numeric matrix containing estimates at each EM iteration.
<code>phi</code>	Converged estimate for tumour ploidy parameter. numeric array containing estimates at each EM iteration.
<code>piG</code>	Converged estimate for initial genotype state distribution. numeric matrix containing estimates at each EM iteration.
<code>piZ</code>	Converged estimate for initial clonal cluster state distribution. numeric matrix containing estimates at each EM iteration.
<code>muR</code>	Mean of binomial mixtures computed as a function of <code>s</code> and <code>n</code> . numeric matrix containing estimates at each EM iteration. See References for mathematical details.
<code>muC</code>	Mean of Gaussian mixtures computed as a function of <code>s</code> , <code>n</code> , and <code>phi</code> . numeric matrix containing estimates at each EM iteration. See References for mathematical details.
<code>loglik</code>	Posterior Log-likelihood that includes data likelihood and the priors. numeric array containing estimates at each EM iteration.
<code>rhoG</code>	Posterior marginal probabilities for the genotype states computed during the E-step. Only the final iteration is returned as a numeric matrix.
<code>rhoZ</code>	Posterior marginal probabilities for the clonal cluster states computed during the E-step. Only the final iteration is returned as a numeric matrix.
<code>genotypeParams</code>	Original genotype parameters. See <code>loadDefaultParameters</code> .
<code>ploidyParams</code>	Original tumour ploidy parameters. See <code>loadDefaultParameters</code> .
<code>normalParams</code>	Original normal contamination parameters. See <code>loadDefaultParameters</code> .
<code>clonalParams</code>	Original subclonal parameters. See <code>loadDefaultParameters</code> .
<code>txnExpLen</code>	Original genotype transition expected length. See <code>loadDefaultParameters</code> .
<code>txnZstrength</code>	Original clonal cluster transition expected length. See <code>loadDefaultParameters</code> .
<code>useOutlierState</code>	Original setting indicating usage of outlier state. See <code>loadDefaultParameters</code> .

**Author(s)**

Gavin Ha <gavinha@gmail.com>

**References**

Ha, G., Roth, A., Khattra, J., Ho, J., Yap, D., Prentice, L. M., Melnyk, N., McPherson, A., Bashashati, A., Laks, E., Biele, J., Ding, J., Le, A., Rosner, J., Shumansky, K., Marra, M. A., Huntsman, D. G., McAlpine, J. N., Aparicio, S. A. J. R., and Shah, S. P. (2014). TITAN: Inference of copy number architectures in clonal cell populations from tumour whole genome sequence data. *Genome Research*, 24: 1881-1893. (PMID: 25060187)

**See Also**

["foreach"](#), ["doMC"](#), ["doMPI"](#), [loadAlleleCounts](#), [loadDefaultParameters](#), [viterbiClonalCN](#)

**Examples**

```
message('Running TITAN ...')
#### LOAD DATA ####
infile <- system.file("extdata", "test_alleleCounts_chr2.txt",
                      package = "TitanCNA")
data <- loadAlleleCounts(infile)

#### LOAD PARAMETERS ####
message('titan: Loading default parameters')
numClusters <- 2
params <- loadDefaultParameters(copyNumber = 5,
                               numberClonalClusters = numClusters, skew = 0.1)

#### READ COPY NUMBER FROM HMMCOPY FILE ####
message('titan: Correcting GC content and mappability biases...')
tumWig <- system.file("extdata", "test_tum_chr2.wig", package = "TitanCNA")
normWig <- system.file("extdata", "test_norm_chr2.wig", package = "TitanCNA")
gc <- system.file("extdata", "gc_chr2.wig", package = "TitanCNA")
map <- system.file("extdata", "map_chr2.wig", package = "TitanCNA")
cnData <- correctReadDepth(tumWig, normWig, gc, map)
logR <- getPositionOverlap(data$chr, data$posn, cnData)
data$logR <- log(2^logR) #transform to natural log

#### FILTER DATA FOR DEPTH, MAPPABILITY, NA, etc ####
data <- filterData(data, 1:24, minDepth = 10, maxDepth = 200, map = NULL)

#### EM (FWD-BACK) TO TRAIN PARAMETERS ####
#### Can use parallelization packages ####
K <- length(params$genotypeParams$alphaKHyper)
params$genotypeParams$alphaKHyper <- rep(500, K)
params$ploidyParams$phi_0 <- 1.5
convergeParams <- runEMclonalCN(data, params,
                               maxiter = 3, maxiterUpdate = 500,
                               txnExpLen = 1e15, txnZstrength = 5e5,
                               useOutlierState = FALSE,
                               normalEstimateMethod = "map",
                               estimates = TRUE, estimatePloidy = TRUE)
```

---

TitanCNA trained dataset

*TITAN EM trained results for an example dataset*

---

### Description

Data for chromosome 2 for a triple-negative breast cancer dataset and the expectation-maximization (EM) trained results. Only 20,000 datapoints are included and the data has been scrambled to anonymize patient SNPs.

**data** Processed input data that is first generated by `loadAlleleCounts`, and includes log ratios that have been GC content and mappability corrected using `correctReadDepth`.

**convergeParams** EM results that are generated by `runEMclonalCN`

### Usage

```
data(EMresults)
```

### Format

'data' is a list. 'convergeParams' is a list.

### References

Shah SP et al. (2012). The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature*, 486(7403): 395-399. (PMID: 22495314)

Ha, G., Roth, A., Lai, D., Bashashati, A., Ding, J., Goya, R., Giuliany, R., Rosner, J., Oloumi, A., Shumansky, K., Chin, S.F., Turashvili, G., Hirst, M., Caldas, C., Marra, M. A., Aparicio, S., and Shah, S. P. (2012). Integrative analysis of genome wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple negative breast cancer. *Genome Research*, 22(10):1995,2007. (PMID: 22637570)

Ha, G., Roth, A., Khattra, J., Ho, J., Yap, D., Prentice, L. M., Melnyk, N., McPherson, A., Bashashati, A., Laks, E., Biele, J., Ding, J., Le, A., Rosner, J., Shumansky, K., Marra, M. A., Huntsman, D. G., McAlpine, J. N., Aparicio, S. A. J. R., and Shah, S. P. (2014). TITAN: Inference of copy number architectures in clonal cell populations from tumour whole genome sequence data. *Genome Research*, 24: 1881-1893. (PMID: 25060187)

---

viterbiClonalCN

*Function to run the Viterbi algorithm for TitanCNA.*

---

### Description

Function to run the Viterbi algorithm to find the optimal state path in the **TitanCNA** hidden Markov model (HMM). The states returned will indicate the optimal copy number and LOH state as well as the most likely clonal cluster for each data point. After running EM, use the converge parameters and the input data to infer the optimal state for each position. This function makes calls to a C subroutine that allows the algorithm to be run more efficiently.

## Usage

```
viterbiClonalCN(data, convergeParams, genotypeParams = NULL)
```

## Arguments

- data** [list](#) object that contains the components for the data to be analyzed. chr, posn, ref, and tumDepth that can be obtained using [loadAlleleCounts](#), and logR that can be obtained using [correctReadDepth](#) and [getPositionOverlap](#) (see Example).
- convergeParams** [list](#) object that is returned from the function [runEMclonalCN](#) in **TitanCNA**.
- genotypeParams** If convergeParams does not contain a genotypeParams element, then the user can pass this as an argument.

## Details

It is difficult to interpret the output of this function directly. The user should use the function [outputTitanResults](#) after.

## Value

numeric [array](#) containing the integer states corresponding to each data point in data.

## Author(s)

Gavin Ha <gavinha@gmail.com>

## References

Ha, G., Roth, A., Khattra, J., Ho, J., Yap, D., Prentice, L. M., Melnyk, N., McPherson, A., Bashashati, A., Laks, E., Biele, J., Ding, J., Le, A., Rosner, J., Shumansky, K., Marra, M. A., Huntsman, D. G., McAlpine, J. N., Aparicio, S. A. J. R., and Shah, S. P. (2014). TITAN: Inference of copy number architectures in clonal cell populations from tumour whole genome sequence data. *Genome Research*, 24: 1881-1893. (PMID: 25060187)

## See Also

[outputTitanResults](#), [loadAlleleCounts](#)

## Examples

```
data(EMresults)

#### COMPUTE OPTIMAL STATE PATH USING VITERBI ####
optimalPath <- viterbiClonalCN(data, convergeParams)
```

**Description**

Fast fixedStep WIG file reading and parsing

**Usage**

```
wigToRangedData(wigfile, verbose = TRUE)
```

**Arguments**

wigfile	Filepath to fixedStep WIG format file
verbose	Set to FALSE to suppress messages

**Details**

Reads the entire file into memory, then processes the file in rapid fashion, thus performance will be limited by memory capacity.

The WIG file is expected to conform to the minimal fixedStep WIG format (see References), where each chromosome is started by a “fixedStep” declaration line. The function assumes only a single track in the WIG file, and will ignore any lines before the first line starting with “fixedStep”.

**Value**

[RangedData](#) for [wigToRangedData](#) with chromosome and position information, sorted in decreasing chromosomal size and increasing position.

**Author(s)**

Daniel Lai

**References**

**WIG** <http://genome.ucsc.edu/goldenPath/help/wiggle.html>

**See Also**

[wigToRangedData](#) is a wrapper around these functions for easy WIG file loading and structure formatting. It is taken from the **HMMcopy** package.

**Examples**

```
map <- system.file("extdata", "map_chr2.wig", package = "TitanCNA")
mScore <- as.data.frame(wigToRangedData(map))
```

# Index

- \*Topic **IO**
  - correctReadDepth, 6
  - extractAlleleReadCounts, 7
  - haplotype-analysis-methods, 12
  - loadAlleleCounts, 14
  - output-methods, 18
  - TitanCNA-package, 2
  - WIG Import Functions, 30
- \*Topic **aplot**
  - Plotting TITAN results, 21
- \*Topic **attribute**
  - loadDefaultParameters, 16
- \*Topic **color**
  - Plotting TITAN results, 21
- \*Topic **datasets**
  - TitanCNA trained dataset, 28
- \*Topic **htest**
  - runEMclonalCN, 24
  - viterbiClonalCN, 28
- \*Topic **iteration**
  - runEMclonalCN, 24
  - viterbiClonalCN, 28
- \*Topic **manip**
  - computeSdbwIndex, 4
  - correctReadDepth, 6
  - filterData, 9
  - getPositionOverlap, 10
  - output-methods, 18
  - runEMclonalCN, 24
  - TitanCNA-package, 2
  - viterbiClonalCN, 28
- \*Topic **models**
  - runEMclonalCN, 24
  - viterbiClonalCN, 28
- \*Topic **multivariate**
  - runEMclonalCN, 24
  - viterbiClonalCN, 28
- \*Topic **package**
  - TitanCNA-package, 2
- array, 11, 18, 29
- computeSdbwIndex, 4, 19, 21, 24
- convergeParams (TitanCNA trained dataset), 28
- correctReadcount, 6, 7
- correctReadDepth, 6, 11, 18, 25, 28, 29
- data (TitanCNA trained dataset), 28
- data.frame, 6, 8, 9, 11, 22
- data.table, 9, 12–15, 20
- EMresults (TitanCNA trained dataset), 28
- extractAlleleReadCounts, 7
- filterData, 9
- foreach, 25, 27
- getHaplotypesFromVCF (haplotype-analysis-methods), 12
- getPositionOverlap, 10, 18, 25, 29
- haplotype-analysis-methods, 12
- list, 5, 13, 17, 18, 20, 25, 26, 29
- loadAlleleCounts, 4, 5, 9–11, 14, 16–18, 25, 27–29
- loadDefaultParameters, 14, 15, 16, 25–27
- loadHaplotypeAlleleCounts, 23
- loadHaplotypeAlleleCounts (haplotype-analysis-methods), 12
- output-methods, 18
- outputModelParameters, 5, 19, 20
- outputModelParameters (output-methods), 18
- outputTitanResults, 4, 18–20, 22, 24, 26, 29
- outputTitanResults (output-methods), 18
- outputTitanSegments (output-methods), 18
- PileupParam, 8
- plot, 23
- plotAllelicRatio, 23
- plotAllelicRatio (Plotting TITAN results), 21
- plotClonalFrequency, 23

plotClonalFrequency (Plotting TITAN results), 21  
plotCNlogRByChr, 22, 23  
plotCNlogRByChr (Plotting TITAN results), 21  
plotHaplotypeFraction, 14, 23  
plotHaplotypeFraction (Plotting TITAN results), 21  
plotSegmentMedians, 23  
plotSegmentMedians (Plotting TITAN results), 21  
plotSubcloneProfiles, 23  
plotSubcloneProfiles (Plotting TITAN results), 21  
Plotting TITAN results, 21

RangedData, 30  
runEMclonalCN, 18, 21, 22, 24, 24, 28, 29

setGenomeStyle, 13  
setGenomeStyle (loadAlleleCounts), 14

TitanCNA (TitanCNA-package), 2  
TitanCNA trained dataset, 28  
TitanCNA-dataset (TitanCNA trained dataset), 28  
TitanCNA-package, 2

viterbiClonalCN, 18, 21, 27, 28

WIG Import Functions, 30  
wigToRangedData, 6, 7, 30  
wigToRangedData (WIG Import Functions), 30