Package 'DBSverify'

August 5, 2021

Title Analyze Next Generation Sequencing Data to Verify Somatic Double-BaseSubstituion (DBS) Calls

Version 0.10

Description DBSverify stands for ``Doublet Base Substitution verify". The main function is Read_DBS_VCF_and_BAMs_to_verify_DBSs. This reads a VCF (variant call format) file containing somatic DBS mutations and examines the supporting reads in the corresponding tumor and normal BAM files to assess wither each DBS is likely real, i.e. arose at one time on a single chromosome. This is as opposed to being two adjacent single base mutations on homologous chromosomes, or adjacent mutations that occurred on the same chromosome but at detectably different times. Many somatic mutation callers do not call DBSs at all, and those that do sometimes make obvious errors, such as calling DBSs that consist of a germline SNP next to a somatic single base mutation. Because of the way the functionality of this package is used, the main function for most users will be Read_DBS_VCF_and_BAMs_to_verify_DBSs.

License GPL-3

Encoding UTF-8

LazyData true

Roxygen list(markdown = TRUE)

RoxygenNote 7.1.1

Imports stringr,
ICAMS (>= 3.0.0),
data.table

Suggests dplyr,
GenomicRanges,
spelling,
testthat,
tibble

Language en-US

R topics documented:

`CategorizeReads` *Categorizes the reads in one sam file.*

Description

Categorizes the reads in one sam file.

Usage

```
CategorizeReads(sam, POS, REF, ALT)
```

Arguments

`sam`	An in-memory version of a sam file as data.frame.
`POS`	The position in the sam at which the DBS starts.
`REF`	The reference allele.
`ALT`	The putative alternative allele.

Value

a named character vector, the names are the read names in the sam file. Each element is one of "Read supports only 1st position" "Read supports only 2nd position", "WT read" "Mut read"

@keywords internal

`DBS_conclusion_1_row`
Examine 1 row of a DBS-only VCF that was already processed by `Slice2ReadSupport` *to decide if the DBS is likely real.*

Description

Examine 1 row of a DBS-only VCF that was already processed by `Slice2ReadSupport` to decide if the DBS is likely real.

Usage

```
DBS_conclusion_1_row(row, germlineCutOff = 0.1, max.half.support.T.reads = 1)
```

Arguments

`row`	One row of a DBS VCF (as a `data.frame`) already processed by `Slice2ReadSupport` (so that the fields `vcf$NreadSupport` and `vcf$TreadSupport` are populated).
`germlineCutOff`	If this proportion of normal reads show one or the other variant (or both variants), consider this a germline variant or partial germline variant.
`max.half.support.T.reads`	Do not tolerate more than this number of reads in the tumor that support one but not both mutated positions.

Value

A character string indicating the conclusion about the putative DBS.

`GetAllBAMSlicesSamtools` *Get and save all slices of BAM as specified by positions in a VCF table.*

Description

Get and save all slices of BAM as specified by positions in a VCF table.

Usage

```
GetAllBAMSlicesSamtools(
  vcf,
  bam.name,
  padding = 10,
  where.to.put.slices = tempfile(),
  verbose = 0
)
```

Arguments

`vcf`	A VCF (Variant Call Format) "file" as a data.table or similar.
`bam.name`	The file name (path) to the BAM file to slice.
`padding`	How many base pairs on each side of the first base of the DBS to to keep in the BAM slices.
`where.to.put.slices`	If `NULL`, create a temporary directory to store the slices. Otherwise, a character string that specifies a directory in which to store the BAM slices. This is directory is created if necessary.
`verbose`	If > 0 print a message when starting the number of slices generated every `verbose` slices.

Value

A character string that specifies the directory the containing the BAM slices. The slices are stored as SAM files.

PCAWG_read_table_and_evaluate_DBS *Read a table that specifies how to process VCFs and mini-BAMS to evaluate DBS calls, for PCAWG Collaboratory data.*

Description

Read a table that specifies how to process VCFs and mini-BAMS to evaluate DBS calls, for PCAWG Collaboratory data.

Usage

```
PCAWG_read_table_and_evaluate_DBS(
  in.table,
  in.vcf.dir,
  minibam.dir,
  out.vcf.dir = in.vcf.dir,
  bam.suffix = "_dbs_srt",
  verbose = 1
)
```

Arguments

in.table	the file path of the table to process; in production, .../DBSverify/data-raw/collaboratory_bams_full_20 for testing .../DBSverify/data-raw/short_collaboratory_bams.csv.
in.vcf.dir	The path to the directory containing the DBS VCF files.
minibam.dir	The path to the directory containing the mini BAMs.
out.vcf.dir	The path to the directory in which to put the "evaluated" DBS VCF files.
bam.suffix	String to add to end of BAM file name; depends on the conventions used by the script (run on the Collaboratory) that generated the miniBAMs.
verbose	If > 0 generate some progress messages.

Details

This is a specialized function for processing PCAWG data from the ICGC (International Cancer Genome) "Collaboratory" cloud computing system, once the miniBAMs have been created in the Collaboratory and downloaded. The in.table and associated BED files were used to specify the contents of the miniBAMs. The result consists of the "evaluated" DBS VCF files. The naming of the input and output VCF files and the mini BAMs is governed by the contents of in.table, with the VCF file names incorporating the "aliquot_id" and the miniBAM names based on the icgc_donor_id and the T_Specimen ID and N_Specimen ID.

`ReadSamfile` *Read a "SAM file", discarding some reads that cannot be interpreted for our purposes.*

Description

Read a "SAM file", discarding some reads that cannot be interpreted for our purposes.

Usage

```
ReadSamfile(filename)
```

Arguments

`filename`	The name of the SAM file to read.

Details

SAM stands for "Sequence Alignment Map", a text file that represents aligned next-generation sequencing reads (https://en.wikipedia.org/wiki/SAM_(file_format)). Only keep reads that satisfy certain conditions:

- Mate pair maps to same chromosome
- Mapping quality >= 30
- "FLAG" < 256 (see info on the return value element `reads.with.bad.FLAG`, below).
- The CIGAR string is only `\d+M` (one or more digits followed M, and nothing else before or after). This means there are no insertions or deletions in the read versus the reference and there is no soft clipping. Insertions or deletions or clipping shift the read within the SAM file, after which this function cannot keep track of the DBS location.

Value

A a list with the elements:

- `good.reads`, `data.frame` with with column names for the first 11 columns as specified in https://en.wikipedia.org/wiki/SAM_(file_format) with one row per read. The result contains only the reads that meet certain conditions.
- `total.depth`, the initial depth including "bad" reads
- `reads.with.bad.FLAG`. These are reads with FLAG >= 256, which marks reads that (i) are "not primary alignment" (ii) failed vendor QC (iii) are PCR or optical duplicates (iv) are supplementary alignments (e.g. split, split /inverted read). See https://broadinstitute.github.io/picard/explain-flags.html
- `reads.with.bad.CIGAR`, reads with CIGAR string that indicates an indel in the read or clipping.
- `reads.with.bad.MAPQ`, reads with MAPQ < 30.
- `reads.with.bad.Mate_CHROM`, reads with a mate on a different chromosome.

`Read_DBS_VCF_and_BAMs_to_verify_DBSs` *Determine whether sequencing reads in fact support (candidate) DBSs present in a VCF file.*

Description

Determine whether sequencing reads in fact support (candidate) DBSs present in a VCF file.

Usage

```
Read_DBS_VCF_and_BAMs_to_verify_DBSs(
  input.vcf,
  Nbam.name,
  Tbam.name,
  N.slice.dir = tempfile(),
  T.slice.dir = tempfile(),
  unlink.slice.dir = TRUE,
  exclude.SBSs = TRUE,
  verbose = 0,
  outfile = NULL,
  filter.status = "PASS"
)
```

Arguments

`input.vcf`	If a character string, then the path to a VCF file; otherwise A a single VCF "file" as a data.frame or similar object.
`Nbam.name`	The name of the BAM file for the normal sample corresponding to `vcf.name`.
`Tbam.name`	The name of the BAM file for the tumor sample corresponding to `vcf.name`.
`N.slice.dir`	Directory for the slices of the normal BAM. Created if necessary.
`T.slice.dir`	Directory for the slices of the tumor BAM. Created if necessary. Must be different than `N.slice.dir`.
`unlink.slice.dir`	If `TRUE` unlink `N.slice.dir` and `T.slice.dir` before return.
`exclude.SBSs`	If `TRUE` silently filter out (exclude) SBSs in the input VCF. This makes sense if the the VCF is from a caller (like Mutect or the Hartwig Medical Foundation caller) that calls both SBSs and DBS.
`verbose`	If > 0 print a message when starting the number of slices generated every `verbose` slices.
`outfile`	If not `NULL` then write the "evaluated" VCF to `outfile`; otherwise write it to `paste0(input.vcf(vcf.name,"_evaluated.vcf")`. Must be non-`NULL` if `input.vcf` is not a file path.
`filter.status`	If not `NULL` only keep rows where the FILTER column in the VCF is equal to `filter.status`.

Details

Creates a new VCF file. This VCF file has no data rows if there were no DBSs to analyze. Otherwise, this VCF contains some additional columns. Any SBSs or indels in the input are silently ignored, and no attempt is made to merge adjacent SBSs.

1. `NreadSupport` With regard to the two positions of the DBS in the normal BAM, a string with 4 numbers separated by ":", with the numbers indicating respectively:
   - the number of reads that are reference sequence at both positions of the DBS,
   - the number of reads that that have the alternative allele only at the 1st position of the DBS,
   - the number of reads that have the alternative allele only at the second position of the DBS, and
   - the number of reads that have the alternative alleles at both positions of the DBS.
2. `TreadSupport` Information analogous to that in `NreadSupport`, for the tumor BAM.
3. `num_bad_mapped_reads` The total number of tumor reads with MAPQ < 30 or with a mate on a different chromosome. If there are many badly mapped reads in the slice the slice may represent a segmental duplication.
4. `num_bad_mapped_DBS_reads` The number of tumor reads with the putative DBS but with MAPQ < 30 or a mate on a different chromosome. If many badly mapped reads support DBSs the DBS might results from mismapped reads in a segmental duplication.
5. `DBSconclusion` A string that describes whether the DBSs is believable (`"True DBS"`), or if the DBS is not believable, a string that describes why not.

The decision in `DBSconclusion` is based on multiple criteria. I also suggest relying on any available upstream filtering of SBSs that get merged into DBSs, as well as upstream filtering of DBSs. It is difficult to capture all the possible characteristics of likely miscalled DBSs, especially if they stem from mismapped reads that nevertheless have high MAPQ (mapping quality). The code that implements these criteria is in `DBS_conclusion_1_row`, which depends on classification of individual reads in `ReadSamfile`. The filtering of reads in the input SAM files is described in the documentation for `ReadSamfile`.

Once the reads to analyze are selected, additional criteria include

- There must be >= 5 normal reads at the site of the putative tumor DBS.
- At each separate position of the DBS, the normal reads must have < 10% of the variant in the DBS at that position.
- < 2 normal reads support the DBS.
- At least 2 tumor reads support the DBS.
- There are more well-mapped tumor reads than badly mapped tumor reads at the site of the DBS.
- There are more well-mapped tumor reads than badly mapped tumor reads that contain the DBS. (`ReadSamfile` keeps track of the well-mapped and badly mapped reads).
- If 1 normal read supports the DBS, then there must be a statistically greater proportion of tumor reads supporting the DBS (by Fisher's test).
- There must be >= 5 normal reads at the site of the putative tumor DBS.
- At each separate position of the DBS, the normal reads must have < 10% of the variant in the DBS at that position.
- < 2 normal reads support the DBS.
- At least 2 tumor reads support the DBS.

- There are more well-mapped tumor reads than badly mapped tumor reads at the site of the DBS.
- There are more well-mapped tumor reads than badly mapped tumor reads that contain the DBS. (`ReadSamfile` keeps track of the well-mapped and badly mapped reads).
- If 1 normal read supports the DBS, then there must be a statistically greater proportion of tumor reads supporting the DBS (by Fisher's test).

Value

Invisibly, a list with the elements

1. The name of the DBS-only VCF file created.
2. The in-memory representation of the DBS VCF as a `data.table`.
3. The name of the directory with the normal SAM slices, if `unlink.slice.dir` is `FALSE`.
4. The name of the directory with the tumor SAM slices, if `unlink.slice.dir` is `FALSE`.

`Read_SBS_VCF_and_BAMs_to_verify_DBSs`
Aggressively merge SBSs into DBSs then determine whether sequencing reads in fact support DBSs.

Description

Aggressively merge SBSs into DBSs then determine whether sequencing reads in fact support DBSs.

Usage

```
Read_SBS_VCF_and_BAMs_to_verify_DBSs(
  input.vcf,
  Nbam.name,
  Tbam.name,
  variant.caller,
  N.slice.dir = tempfile(),
  T.slice.dir = tempfile(),
  unlink.slice.dir = TRUE,
  outfile = NULL,
  verbose = 0
)
```

Arguments

`input.vcf`	If a character string, then the path to a VCF file; otherwise A a single VCF "file" as a data.frame or similar object.
`Nbam.name`	The name of the BAM file for the normal sample corresponding to `vcf.name`.
`Tbam.name`	The name of the BAM file for the tumor sample corresponding to `vcf.name`.
`variant.caller`	One of `"strelka"`, `"PCAWG"`, or `"unknown"`. Merging adjacent SBS is done by `ReadAndSplitVCFs` in the ICAMS package.

`N.slice.dir`	Directory for the slices of the normal BAM. Created if necessary.
`T.slice.dir`	Directory for the slices of the tumor BAM. Created if necessary. Must be different than `N.slice.dir`.
`unlink.slice.dir`	If `TRUE` unlink `N.slice.dir` and `T.slice.dir` before return.
`outfile`	If not `NULL` then write the "evaluated" VCF to `outfile`; otherwise write it to `paste0(input.vcf(vcf.name,"_evaluated.vcf")`. Must be non-`NULL` if `input.vcf` is not a file path.
`verbose`	If > 0 print a message when starting the number of slices generated every `verbose` slices.

Details

Note: argument `input.vcf` must be a file path. This function creates a new VCF file. See `Read_DBS_VCF_and_BAMs_to_verify_DBSs` for details.

Value

Same as `Read_DBS_VCF_and_BAMs_to_verify_DBSs`.

`Slice2ReadSupport` *Calculate the support for a putative DBS from the SAM slice containing the overlapping reads.*

Description

Calculate the support for a putative DBS from the SAM slice containing the overlapping reads.

Usage

```
Slice2ReadSupport(slice.dir, CHROM, POS, REF, ALT)
```

Arguments

`slice.dir`	The directory containing the SAM slices.
`CHROM`	The chromosome identifier.
`POS`	The first position of the DBS.
`REF`	The reference variant.
`ALT`	The alternate variant.

Index