Editing Index Documentation

1 ALIGNMENT

The alignment is recommended to be unique, even though the program filters out any non-primary alignments. This is because some algorithms may generate different "primary" alignment when mapping without unique mapping restriction than with it, which may alter the results in unpredicted ways.

2 REFERENCE GENOME ISSUES

2.1 SKIPPED POSITIONS IN THE REFERENCE GENOME

'N's in the reference sequence of the region will result in skipping of these positions (i.e. these positions will be ignored both for coverage and mismatches). SNPs (as provided by the SNPs table) are skipped as well.

3 OUTPUT DESCRIPTION

3.1 CMPILEUP FILES

These are intermediate files which are pileup files converted to a numerical tabular form. These include however only the coordinate and not the reference bases (saves a lot of runtime). The columns are <region chromosome>, <region start>, <region end>, <position in region>, <always N>, <total coverage>, <#adenosines>, <#cytosines>, <#guanosines>, <#thymines>, <#unrecognized readings>, <# reading under the defined threshold (default is 30)>. These are deleted after processing unless the downstream processing has failed or if specified so by --keep_cmpileup.

3.2 STRANDDECIDINGMETHOD

For non-stranded data, the original strand of the RNA molecule (fixed to be the predicted mainly expressed) must be determined. The following options differ by how the strand is decided (per region, per sample):

- **(Use only this option whenever good gene annotations are available)** ***RefSeqThenMMSites*** - The strand is determined according to gene annotations (i.e. RefSeq annotations), by choosing the strand which is more likely expressed and sequenced (exon>>intron>>intergenic, try to solve conflicts according to known gene expression levels). If the annotation is unavailable or indecisive (such as exons on both strand), strand is decided according to the strand that fits the majority mismatch sites (e.g. if the number of A-to-G mismatch sites is higher- choose sense strand). **This is the cleanest method and the only one present in the non-verbose output.**
- (verbose only) ***MMSitesThenRefSeq*** – The decision is made in the exact opposite order of the previous one – decide according the strand fitting the majority of the mismatch sites, and only if this is indecisive use annotations. Since this method chooses strand according to the mismatches sites, it maximizes noise as well, thus it is not recommended for general usage. One can estimate though how good the annotations were when comparing to the SNR of this output.

- (verbose only) ***Randomly*** – This is mainly a negative control. Ideally, signal would be half as much, noise still the same.

For stranded input, the strand is the strand as decided by the sequencing and alignment. (StrandDecidingMethod in verbose mode is "StrandedData").

3.3 SAMPLE METADATA

These include the ***Sample***, ***SamplePath***, and ***Group*** columns. This is optional input used to associate groups with the samples in the output file (saving merging in downstream analyses).

3.4 INDEX VALUES

For each type of the (non-stranded or stranded depending on input) mismatches, ***<mismatch type>EditingIndex***, represents the index when calculated for that type of mismatch. Normally, A2GEditingindex is the only "real" index, whereas the other represent noise, the (usually) most prominent of which is the ***C2TEditingIndex*** which is of the most common genomic mutation.

3.5 COVERAGE AND MISMATCHES DATA (VERBOSE ONLY)

For each type of base, counts of (relative to sense strand):

1. ***NumOf<base type>PositionsCovered*** **-** Number of reference genome positions of it indexed
2. ***TotalCoverageAt<base type>Positions*** **-** Coverage at these positions (including mismatches, without SNPs)

For each type of mismatch:

1. ***NumOf<mismatch type>Mismatches*** -The number of mismatched bases (per mismatch type)
2. ***NumOf<mismatch type>MismatchesSites*** -The number of mismatched bases (per mismatch type)
3. Both of the above at the given SNPs.

3.6 INDEXED DATA COUNTS (VERBOSE ONLY)

For each type of the non-stranded mismatches (data is not unique count, e.g. if position A was covered by 10 read and B by 5 total coverage would be 15 not 2):

1. The number of indexed "canonical" bases - ***IndexedCanonicalOf<mismatch type>***
2. The number of indexed mismtched bases – ***IndexedMismatchesOf<mismatch type>***
3. The number of mismatches sites indexed over – ***NumOfIndexedMismatchesSitesOf<mismatch type>***
4. The number of total sites indexed over – ***NumOfIndexedOverallSitesOf<mismatch type>***
5. The number of regions with coverage - ***NumOfRegionsWithCoverageFor<mismatch type>***

3.7 STRAND AND REGION TYPE DATA (VERBOSE ONLY)

For each of the non-stranded mismatch type, same data as the previously described but divided to +, -, and unknown strands, and for each of the RefSeq annotations type (exonic, intronic, intergenic).

4 INPUT FILES

4.1 USER SUBMITTED DATA FILES

In all user submitted files, the formats are as found in the UCSC Genome Browser, no headers are allowed (or they can't be intersected), and the files can either be uncompressed or compressed using gzip.

4.1.1 Annotation and Gene Info

These are used for filtrations and strand determination of editing in the data.

4.1.1.1 SNPs

SNPs must include only genomic SNPs. Should be provided in BED format with the following columns and in this order:

1. Chromosome
2. Start (0-based)
3. End
4. Strand
5. <Reference sequence>
6. <SNP sequences> (separated by '/', '-' means no bases, as in insertions and deletions)
7. Genomic function (e.g. intergenic, intronic)
8. Alleles frequencies <Reference>/<SNP>

4.1.1.2 RefSeq Annotations

Should be provided in BED format with the following columns and in this order:

1. Chromosome
2. Start (0-based)
3. End
4. RefSeq Name (e.g. NR_075077.1)
5. Gene Common name (e.g. C1orf141)
6. Strand
7. Exons Start Positions (0-based) separated by ","
8. Exons End Positions separated by ","

4.1.1.3 Average Gene Expression Levels

Should be provided in BED format with the following columns and in this order:

1. Chromosome
2. Start (0-based)
3. End
4. Gene Common name (e.g. C1orf141)
5. Average RFKPMs (each average refers to a single tissue) separated by ","

6. Strand

4.1.2 Edited Regions File

This is the file containing the regions in which the index is calculated. It should be provided in BED format, with the following columns and in this order:

1. Chromosome
2. Start (0-based)
3. End

Note: regions data is by definition not stranded (as editing may occur in both strands).

4.1.3 Genomes

Genomes should be provided in the FASTA format and uncompressed.

4.1.4 Samples Data

4.1.4.1 BAM Files

BAM files names should be in the format of <Sample Name><BAM Suffix>. Alignments should be unique, sorting is not mandatory.

4.1.4.2 Groups and Samples File

This file is used to determine the metadata of each sample. Should be in CSV format and contain these columns (named exactly like here, order is not significant):

1. *Sample* – the sample's name.
2. *SamplePath* – an injective path to the sample's file(s) and should end with '*' (used instead of mate number in cases of paired end data).
3. *Group* – The group this sample is associated to. A sample cannot be associated with more than one group.
4. Any other headers are ignored.

5 USAGE DOCUMENTATION

5.1 COMMAND LINE PARAMETERS

5.1.1 Input options

5.1.1.1 Path options

5.1.1.1.1 *-d, --root_dir* The input directory.

-r, --recursion_depth The depth of recursion into the folders looking for the BAM files.

-x, --excluded_prefixes A list of strings that if are found in a path of a BAM file it won't be included.

-s, --subdirs_prefixes A list of strings that must be found in a path of a BAM file for it to be included.

--exclude_operator The operator used with the prefixes from *excluded* (Or or And)

--include_operator The operator used with the prefixes from *subdirs_prefixes* (Or or And)

--follow_links A flag. If set, the recursive dir will follow symlinks in the input directory.

5.1.1.1.1.1 *-f ,--bam_files_suffix* The suffix of the BAM files to run on (e.g. <Sample Name>Sorted.By.Coord.bam). Should be the full suffix (i.e. everything that comes after the sample name)

5.1.1.2 Input Files Processing

--stranded If set will treat input as stranded, choosing strand using a majority vote RefSeq annotations per "mate" file.

--is_paired_end If set will treat input as paired-end sequencing, this currently affects only PE stranded libraries.

5.1.1.3 Info and Genomic Files

5.1.1.3.1 Samples Info

-g, --groups_file A path to an groups and samples csv file containing their data as defined.

5.1.1.3.2 Data Resources to Use

--genome A default **set of data files** to use (currently for hg19, mm9, hg38, mm10; using Alu regions for human data and B1 and B2 for murine data). Use the next options to override any of them separately. Default resources files are specified in the resource file. If "UserProvided" is chosen no data file are assumed to be built-in.

5.1.1.3.3 Genome for Pileup

-gf, --genome The path to the genome FASTA to use instead of the built-in sets.

5.1.1.3.4 Edited Regions Coordinates

-rb, --regions The path to the edited regions bed file to use instead of the built-in sets.

5.1.1.3.5 Genes Info and SNPs

--snps A path to a SNPs file to use instead of the built-in names.

--refseq A path to a RefSeq annotations file to use instead of the built-in sets.

--genes_expression A path to a genes expression file to use instead of the built-in sets.

5.1.1.4 Configuration Options

-c , --config_override_file A path to a configuration file (in INI format) to override the values in the editing index configuration file found in <install_path>/src/RNAEditingIndex/Configs.

-a, --args Named arguments (in the format of <var>=\"<val>\" <var>=\"<val>\") to override in the defaults configuration. This option is meant to be used for parameters needed to be changed frequently, instead of creating an overriding configuration for each small change.

5.1.2 Output Options

-o, --output_dir The root directory for the cmpileups outputs and other temporary files. Outputs (will create sub-directories per sample according to their original directories tree from the root input directory to the BAM file). Files are deleted at the end of the run (except for cmpileups if otherwise specified)

-os ,--output_dir_summery The directory for the summary outputs. (The script will create sub-directories per sample according to their original directories tree from the root input directory to the BAM file).

5.1.2.1.1.1 *-l , --log_path* The path where the logs will be written.

5.1.2.1.1.2 *--keep_cmpileup* If set will not delete CMPileups

5.1.2.1.1.3 *--verbose* If set output the verbose output format

5.1.3 System Options

--ts The number of threads to use when processing samples (i.e. how many sample to process in parallel). This affects only post-pileup processing.

--tsd The maximal number of threads for strand decisions per sample to use when processing samples.

5.1.4 Stranded Options

--stranded If set, will treat the data as stranded

--paired_end If set, will treat the data as stranded paired-end (Omit this flag to run single-end stranded analysis).

5.2 CONFIGURATION FILES

The configuration overriding options are for overriding options in the DefaultsConfig.ini file. These include mainly name formats (in python INI format) but also some other pileup processing options. Please check inside the configuration file for more info.

6 SYSTEM REQUIREMENTS & INSTALLATION

All requirements and installation guide are detailed in the README file.