SMORE

Synteny Modulator Of Repetitive Elements

Bioinformatics Leipzig

Sarah J. Berkemer, Anne Hoffmann, Cameron R. Murray, Peter F. Stadler

11.09.2017

Contents

1 Introduction **3**

2 Quickstart **3**

2.1 Software Requirements 4

2.2 Mode Overview 4

3 Input **4**

3.1 General input 5

3.2 Gene List 5

3.3 Infernal 6

3.4 tRNAscan-SE 6

3.5 Loci List 7

4 Output **8**

4.1 General Output 8

4.2 Additional files 9

4.3 Verbose Output 9

4.4 Output `smore mix` 10

4.5 Output `smore roast` 10

5 Subcommands **10**

5.1 `smore bake` 10

5.2 `smore prep` 11

5.3 `smore toast` 12

5.4 `smore mix` 14

5.5 `smore roast` 16

5.6 `smore eat` 18

5.7 Summary of Subcommands 19

6 Example **19**

7 Theory **19**

8 Help Pages **19**

9 Contact **19**

1 Introduction

SMORE is a pipeline to detect evolutionary events along a phylogenetic tree based on a set of repetitive elements and multiple sequence alignments. This manual provides a more detailed description of the SMORE pipeline published in [1] and [2].
An example is shown in Section 6.

2 Quickstart

The program can be downloaded here https://github.com/AnneHoffmann/Smore. No installation is required to run the SMORE pipeline except when used with additional programs such as infernal or tRNAscan-SE (see the following subsection for details).

A usual pipeline run consists of two steps: i) retrieving the clusters of genetic elements from the input data and ii) analysing the clusters and counting evolutionary events.

Parts of the pipeline can be used independently from each other in order to repeat some processing steps. Please see Section 5 (Subcommands) for further details.

To start a complete run of the SMORE pipeline, use

```
smore bake --out OUTPATH --ref REFERENCE_SPECIES
            --maf PATH_TO_MAF_FILES --genes GENELIST --newick NEWICK_TREE
```

Upper case terms in the command need to be replaced by the specific information needed. Maf files are multiple sequence alignments in MultiZ format. All maf files should be loacted in the same folder. Newick tree is a tree in newick format. Only species that occur in the tree can be analysed. Each maf alignment is based on a reference species. Please give the species identifier of the maf reference species in the –ref option. The genelist is a table of genetic elements that should be analysed using SMORE. The format is a tab-separated table with the following columns (please write 'NA' for missing data):

```
chr start_pos end_pos species strand type pseudogene seq struc comment
```

Note that species identifier have to be the same in the maf files, gene list, newick tree and reference species!

The output will be written to the specified output directory (--out option).
Instead of using a gene list as input, it is possible to use tRNAscan-SE for the detection

of tRNAs or to use `infernal` together with a covariance model(CM) for the automatic detection of genetic elements specified by the CM. For a more detailed description, see Section 3. The output contains a list of genetic clusters as well as a phylogenetic tree where nodes are labeled with number of evolutionary events. For more detailed information, see Section 4.

2.1 Software Requirements

`SMORE` is programmed in `python` and `perl`, hence it requires python 3 and perl 5. Additionally, it requires a current version of `infernal` [3] and `tRNAscan-SE` [4] in case genetic elements are automatically detected using a covariance model or using `tRNAscan-SE`.

2.2 Mode Overview

The `SMORE` pipeline includes several subcommands, that can be used independently depending on the current data set. Additionally, the subcommands `smore prep` and `smore bake` can be applied using different modes of preprocessing. Fig. 1 gives an overview of available modes. The modes will be explained in detail in the next section, Sec. 3 which will used the subcommand `smore bake` in order to explain the modes. All other subcommands will be explained in a later section, see Sec. 5. Help pages for the pipeline can be called using `smore --help` or directly for each subcommand using `smore [subcommand] --help`.

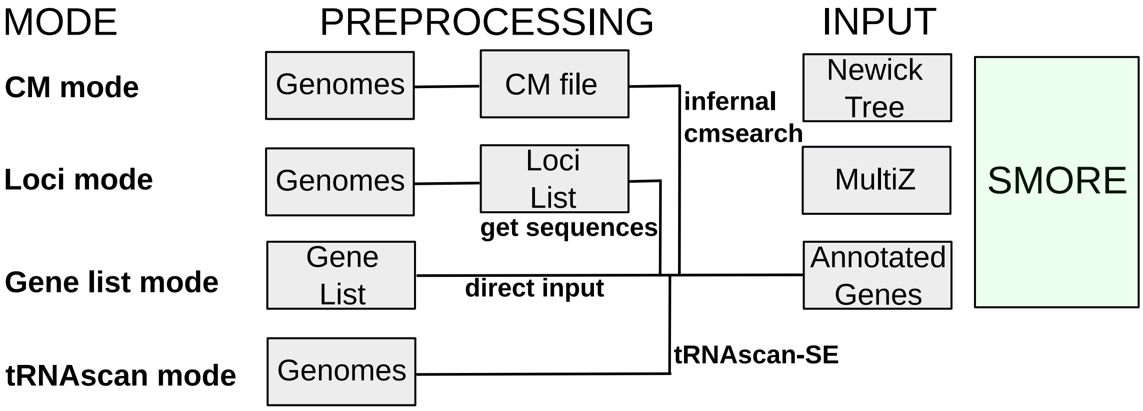

Figure 1: Overview of preprocessing modes of the `SMORE` pipeline.

3 Input

`smore bake` can be used running different modes of how the input gene list is created, hence different sets of parameters are required for different modes of running. This section lists all the parameters sorted by usage mode.

3.1 General input

The following parameters are used in all modes:

option	description
Obligatory parameter	
`--out|-o OUTPATH`	directory where to write the output files. If the folder does not exist, it will be created.
Optional parameter	
`--tool|-t TOOLPATH`	directory where *SMORE* is located. This parameter can usually be omitted.
`--python PYTHON_PATH`	Path to `python` installation if the one specified in the path environment is not used.
`--perl PERL_PATH`	Path to `perl` installation if the one specified in the path environment is not used.
`--filter NUM`	percentage of low scoring maf blocks to be discarded, NUM is between 0 and 1. This parameter is optional, default = 0.

3.2 Gene List

option	description
`--genes GENELIST`	a list of genetic elements as input.

The gene list is a tab-separated table with one entry for each gene. The format is the following:

```
chromosome start_pos end_pos species strand type pseudogene sequence structure
comment
```

The following constraints apply:

- fields to be omitted or optional fields that are not used are filled with NA

- species name should be the same as in MultiZ files
- type is optional (e.g. tRNA type Met)
- pseudogene (optional): true or false
- structure is optional, dot-bracket-notation
- for optional sequence, take loci mode
- comment is optional

Note that species identifier and chromosome identifier have to be the same in the maf files, gene list, newick tree and reference species!

3.3 Infernal

option	description
`--cm|-c CM`	covariance model file, input for infernal
`--genomes|-g GENOMES FOLDER`	folder with genomes of the species used as an input to infernal to scan the genomes for genetic elements specified by the CM. Filenames should match species names in MultiZ files.
`--incE NUM`	optional parameter for cmsearch (infernal), e-value threshold.
`--incT NUM`	optional parameter for cmsearch (infernal), bitscore threshold.
`--infernal PATH`	path to cmsearch, only needed if the one specified in the path environment is not used.
`--pseudo NUM`	optional parameter, bitscore threshold that defines pseudogenes.

3.4 tRNAscan-SE

option	description
`--genomes|-g GENOMES FOLDER`	folder with genomes of the species used as an input to infernal to scan the genomes for genetic elements specified by the CM. Filenames should match species names in MultiZ files.
`--trna`	option to activate the usage of tRNAscan-SE on the given genomes.
`--trnascan PATH`	path to tRNAscan-SE, only needed if it is not installed in the environment path variable.

3.5 Loci List

The loci list mode can be applied if only coordinates of elements are known. Given the corresponding genomes, SMORE will automatically retrieve the gene sequences that are needed for the analysis.

option	description
`--genomes|-g GENOMES FOLDER`	folder with genomes of the species used to retrieve the sequence of the genetic elements in the loci list based on their coordinates. Filenames should match species names in loci list files.
`--loci FILE`	list with genetic elements without sequence

The loci list is a tab-separated table with one entry for each gene. The format is the following:

```
chromosome start_pos end_pos species strand type pseudogene
```

The following constraints apply:

- fields to be omitted or optional fields that are not used are filled with NA
- species name should be the same as in MultiZ files
- type is optional (e.g. tRNA type Met)
- pseudogene (optional): true or false

Note that species identifier and chromosome identifier have to be the same in the maf files, loci list, newick tree and reference species!

4 Output

All output files are located in the specified folder, given with option `--out` or `-o`. The output consists of multiple files that will be described in the following subsections. summary file

4.1 General Output

output file	description
OutTree.txt	the resulting tree in newick format with numbers at the nodes given in brackets. This format can be used to visualize the tree with newick compatible programs.
geneticEvents.txt	File listing all genetic events counted during the analysis. The numbers are sorted by event and node of the tree. The file includes a event called 'Other'. This will give the difference of genetic elements between the total amount and the elements used in the analysis. For a successful run of the pipeline, the numbers should be 0.
data_iTOL	folder giving files that can be uploaded to *itol.embl.de* [5]. The file called F0tree.txt is uploaded at http://itol.embl.de/upload.cgi. The remaining files can be added using drag and drop into the browser window. This will result in a interactive visualization of the resulting tree. A legend is added automatically. All nodes in the tree will have unique names, thus some nodes might have names such as 'innerNode0' because it was added automatically.
allClusters_original.txt, allClusters_joined.txt	These two files contain lists of clusters showing which elements are contained together in one cluster before and after joining.

errors	for each part of the smore pipeline, there is a file giving errors that happened during the run. If no errors occured, the file is empty.
list_cographs.txt, *list_noncographs.txt*	these files contain statistics about graphs that were cographs from the beginning or had to be edited in order to become a cograph. The tables list number of nodes, number of edges and density of the graphs.
remoldings.txt, inremoldings.txt	These files contain genetic elements that (a) have highly similar sequences but different types or (b) have the same types but clearly distinct sequences.
allTypes.txt, allPseudoTypes.txt	These two files list the different types of genetic elements for all species and functional or pseudogenized genes. This can be used to analyse the distribution of different types of genetic elements.

4.2 Additional files

- for each species, there are files listing their genetic elements as they were used as intermediate files.
- there are files for each of the evolutionary events where counts are listed.
- for singletons, there is also a listing about types and pseudogenes
- for elements which could not be sorted in between genomic anchors (nones) there is listings about types and pseudogenes.

4.3 Verbose Output

the verbose version of toast will output three additional files (if `--verbose`) or any combination of them (if `--clus, --graph, --aln`) for each cluster, named with left and right anchor numbers to match all three files. They will be in three different folders: cluster, graph and duplication_alignment. The files contain the specific structures of the cluster in each step of the analysis and can be used to gain a deeper insight into the data.

4.4 Output `smore mix`

4.5 Output `smore roast`

5 Subcommands

In order to repeat only some parts of the pipeline or have a more detailed look into intermediary files, parts of the pipeline can be used independently. This section gives an overview of all possible subcommands explaining input, output and options. Fig. 2 gives a graphical overview of subcommands in comparison to the workflow.

The `SMORE` pipeline can be applied in one complete run but also splitted in several parts that allow the user to change and compare results created by different parameter sets. Additionally, the pipeline of subcommnands consisting of `smore prep`, `smore mix`, `smore roast`, `smore eat` is able to process large amounts of data by splitting the set of clusters in disjoint subsets (`smore mix`), analyse each part individually (`smore roast`) and summarizing all results at the end (`smore eat`).

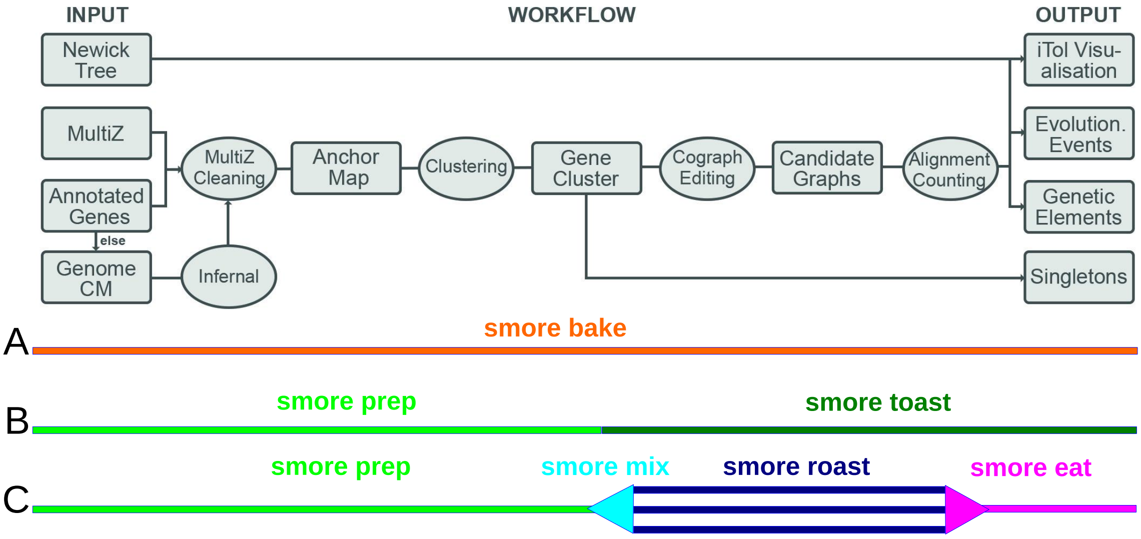

Figure 2: Overview of subcommands of the `SMORE` pipeline next to the workflow.

5.1 `smore bake`

The subcommand `smore bake` will run the complete pipeline for a given input. Different modes such as gene list or infernal mode can be specified by the set of parameters and input data given. The output is saved in the specified output folder (option `--out`). In case the output folder does not exist, it will be created. As `smore bake` is a combination of the subcommands `prep` and `toast`, the parameter set is the same as for `prep` and `toast`.

Thus, please see the tables in the next two subsections for parameter descriptions. The help pages for `smore bake` is `smore bake --help` or `smore bake -h`.

5.2 smore prep

By applying `smore prep` only the first part of the pipeline will be executed. Hence, only input data concerning the creation of genetic clusters is required. This subcommand can be used to create a basic dataset that will be the input of following steps of the pipeline. Afterwards, succeeding subcommands can be applied with different parameter sets. Depending on the number of genetic elements and genome size, `smore prep` might take several hours to process the input data. The output will be saved in the specified output folder (option `--out`). The help page for `smore prep` is `smore prep --help` or `smore prep -h`.

parameter	obl/opt	description
General		
`--out|-o OUTPATH`	obl	directory where to write the output files. If the folder does not exist, it will be created.
`--maf MAF_FILES`	obl	directory where all multiZ alignment files are located.
`--ref REFERENCE_SPECIES`	obl	name of the reference species as specified in the multiZ alignment files.
`--tool|-t TOOLPATH`	opt	directory where *SMORE* is located. This parameter can usually be omitted.
`--python PYTHON_PATH`	opt	Path to `python` installation if the one specified in the path environment is not used.
`--perl PERL_PATH`	opt	Path to `perl` installation if the one specified in the path environment is not used.
`--filter NUM`	opt	percentage of low scoring maf blocks to be discarded, NUM is between 0 and 1. This parameter is optional, default = 0.
Genelist Mode		
`--genes GENELIST`	obl	a list of genetic elements as input.
Infernal mode		
`--cm|-c CM`	obl	covariance model file, input for infernal

`--genomes|-g`
`GENOMES FOLDER`
	obl	folder with genomes of the species used as an input to infernal to scan the genomes for genetic elements specified by the CM. Filenames should match species names in MultiZ files.

`--incE NUM`	opt	parameter for cmsearch (infernal), e-value threshold.
`--incT NUM`	opt	parameter for cmsearch (infernal), bitscore threshold.
`--infernal PATH`	opt	path to cmsearch, only needed if the one specified in the path environment is not used.
`--pseudo NUM`	opt	bitscore threshold that defines pseudogenes.
tRNAscan-SE mode		

`--genomes|-g`
`GENOMES FOLDER`
	obl	folder with genomes of the species used as an input to infernal to scan the genomes for genetic elements specified by the CM. Filenames should match species names in MultiZ files.

`--trna`	obl	option to activate the usage of tRNAscan-SE on the given genomes.
`--trnascan PATH`	opt	path to tRNAscan-SE, only needed if it is not installed in the environment path variable.
Locilist mode		

`--genomes|-g`
`GENOMES FOLDER`
	obl	folder with genomes of the species used to retrieve the sequence of the genetic elements in the loci list based on their coordinates. Filenames should match species names in loci list files.

`--loci FILE`	obl	list with genetic elements without sequence

5.3 `smore toast`

The subcommand `smore toast` can be applied after `smore prep`. It will take `smore prep` output (option `--prep`) and analyse the given genetic clusters. The output will be saved in the specified output folder (option `--out`). As `smore toast` can be applied using different sets of parameters, applying it several times on the same input data can help estimating the best parameter set e.g. joining of clusters (`--join`) or similarity thresholds (`--seq`, `--struc`). By omitting verbose output, `smore toast` is able to process given input data in only several minutes. Verbose output (if `--verbose`) or any combination of them (if

`--clus, --graph, --aln`) will help to get a deeper insight into intermediary steps and how the clusters are composed but will take a lot more time to process the data and more space capacities as all intermediary files are stored in the output folder. The help page for `smore toast` is `smore toast --help`.

parameter	obl/opt	description
`--out|-o OUTPATH`	obl	directory where to write the output files. If the folder does not exist, it will be created.
`--maf MAF_FILES`	obl	directory where all multiZ alignment files are located.
`--ref REFERENCE_SPECIES`	obl	name of the reference species as specified in the multiZ alignment files.
`--newick NEWICK_FILE`	obl	phylogenetic tree in newick format including all species contained in the analysis. For automatic conversion of species identifiers, use parameter –id.
`--id ID_FILE`	opt	in case species identifier in the tree and multiZ files do not match, this file can be used to automatically convert species identifier. Format of the table: current_name_in_tree (tab) name_to_translate_to
`--join LEVEL`	opt	level of how to join the clusters. Levels are *none, strict* and *relaxed.* Default: *relaxed.* The level refers to the adjacency constraints of gene clusters, see Sec. 7 for further details.
`--seqsim NUM`	opt	percentage of sequence similarity in order to define two sequences as orthologs. Number between 0 and 1, default: 0.9
`--strucsim NUM`	opt	percentage of structure similarity in order to define two structures as orthologs. Number between 0 and 1, default: 0.9
`--nomiss`	opt	using this option, the pipeline will omit the check for missing data. Hence, deletions in the tree won't distinguish between a deletion because of missing data or a 'real' deletion. Only useful when running on a large data set.

`--verbose`	opt	using this option, the pipeline will print all intermediary files such as clusters, graphs and alignments. This will slow down the running time of the program but will give deeper insights into the data.
`--clus`	opt	using this option, the pipeline will print all intermediary cluster files
`--graph`	opt	using this option, the pipeline will print all intermediary graph files
`--aln`	opt	using this option, the pipeline will print all intermediary alignment files
`--tool|-t TOOLPATH`	opt	directory where *SMORE* is located. This parameter can usually be omitted.
`--python PYTHON_PATH`	opt	Path to `python` installation if the one specified in the path environment is not used.
`--perl PERL_PATH`	opt	Path to `perl` installation if the one specified in the path environment is not used.

5.4 `smore mix`

`smore mix` is designed to only run the joining of the clusters based on given parameters and `smore prep` output. It will return a list of the original clusters and another file listing clusters after joining. Additionally, `smore mix` can be used to handle large amounts of data. With option `--max` the maximal number of clusters can be specified that should be included in the pipeline's next processing step. `smore mix` will hence split the resulting clusters in several files that do not exceed the specified number. In order to help the user proceed to the next step of the pipeline (`smore roast`), the output will contain a command list, specifying the commands needed to call the next step with each part of the cluster list. The user can easily copy the commands and call the next step or modify the commands by adding parameters. Verbose output can be used with `smore mix` but might slow down running time of the program. The possible parameters for `smore mix` include parameters for `smore roast`, too, as the parameters will be used to create the commandlist. The help page for `smore mix` is `smore mix --help`.

parameter	obl/opt	description
`--out|-o OUTPATH`	obl	directory where to write the output files. If the folder does not exist, it will be created.
`--prep PATH`	obl	directory where `smore prep` output files are located.
`--species FILE`	obl	file listing all species included in the analysis. Thus, the file includes the species names as given in the newick tree and prep output, one species identifier per line.
`--newick NEWICK_FILE`	obl	phylogenetic tree in newick format including all species contained in the analysis. For automatic conversion of species identifiers, use parameter –id.
`--id ID_FILE`	opt	in case species identifier in the tree and multiZ files do not match, this file can be used to automatically convert species identifier. Format of the table: current_name_in_tree (tab) name_to_translate_to
`--join LEVEL`	opt	level of how to join the clusters. Levels are *none, strict* and *relaxed*. Default: *relaxed*. The level refers to the adjacency constraints of gene clusters, see Sec. 7 for further details.
`--max NUM`	opt	this parameter can be used to specify the maximal number of clusters used in following steps of the analysis. If there are more clusters, the program will automatically split the data set and create a command list in order to run the next steps in parallel. The default value is 50000.
`--seqsim NUM`	opt	percentage of sequence similarity in order to define two sequences as orthologs. Number between 0 and 1, default: 0.9
`--strucsim NUM`	opt	percentage of structure similarity in order to define two structures as orthologs. Number between 0 and 1, default: 0.9

<table>
<tr><td><code>--nomiss</code></td><td>opt</td><td>using this option, the pipeline will omit the check for missing data. Hence, deletions in the tree won't distinguish between a deletion because of missing data or a 'real' deletion. Only useful when running on a large data set.</td></tr>
<tr><td><code>--verbose</code></td><td>opt</td><td>using this option, the pipeline will print all intermediary files such as clusters, graphs and alignments. This will slow down the running time of the program but will give deeper insights into the data.</td></tr>
<tr><td><code>--clus</code></td><td>opt</td><td>using this option, the pipeline will print all intermediary cluster files</td></tr>
<tr><td><code>--tool|-t TOOLPATH</code></td><td>opt</td><td>directory where <i>SMORE</i> is located. This parameter can usually be omitted.</td></tr>
<tr><td><code>--python PYTHON_PATH</code></td><td>opt</td><td>Path to <code>python</code> installation if the one specified in the path environment is not used.</td></tr>
<tr><td><code>--perl PERL_PATH</code></td><td>opt</td><td>Path to <code>perl</code> installation if the one specified in the path environment is not used.</td></tr>
</table>

5.5 `smore roast`

The subcommand `smore roast` is applied on the `smore mix` output. The program `smore mix` will output a list of commands on how to proceed. Thus `smore roast` is called based on these commands. However the user can change or add some parameters to the preprinted commands in order to adjust the program call. If there is more than one commands, it is important that all commands are executed with the same parameter set in order to obtain a homogenuous result at the end. If some of the commands are omitted, the final result will only include the data that was analysed. The next step of the pipeline `smore eat` will summarize the `smore roast` output into final results of the pipeline run. The resulting files of `smore roast` are given specified names to make sure that the next step of the pipeline will find all input data. Thus, file names should not be changed and files not be moved to another folder. Verbose output can be used with `smore roast`, also for just some of the commands but might slow down the running time of the program. The help page for `smore roast` is `smore roast --help`.

parameter	obl/opt	description
`--out|-o OUTPATH`	obl	directory where to write the output files. If the folder does not exist, it will be created.
`--in|-i FILE`	obl	Output file of `smore mix` including the list of clusters.
`--species FILE`	obl	file listing all species included in the analysis. Thus, the file includes the species names as given in the newick tree and prep output, one species identifier per line.
`--newick NEWICK_FILE`	obl	phylogenetic tree in newick format including all species contained in the analysis. For automatic conversion of species identifiers, use parameter –id.
`--id ID_FILE`	opt	in case species identifier in the tree and multiZ files do not match, this file can be used to automatically convert species identifier. Format of the table: current_name_in_tree (tab) name_to_translate_to
`--seqsim NUM`	opt	percentage of sequence similarity in order to define two sequences as orthologs. Number between 0 and 1, default: 0.9
`--strucsim NUM`	opt	percentage of structure similarity in order to define two structures as orthologs. Number between 0 and 1, default: 0.9
`--nomiss`	opt	using this option, the pipeline will omit the check for missing data. Hence, deletions in the tree won't distinguish between a deletion because of missing data or a 'real' deletion. Only useful when running on a large data set.
`--verbose`	opt	using this option, the pipeline will print all intermediary files such as clusters, graphs and alignments. This will slow down the running time of the program but will give deeper insights into the data.
`--graph`	opt	using this option, the pipeline will print all intermediary graph files
`--aln`	opt	using this option, the pipeline will print all intermediary duplication alignment files

`--tool|-t TOOLPATH`	opt	directory where *SMORE* is located. This parameter can usually be omitted.
`--python PYTHON_PATH`	opt	Path to `python` installation if the one specified in the path environment is not used.
`--perl PERL_PATH`	opt	Path to `perl` installation if the one specified in the path environment is not used.

5.6 `smore eat`

The subcommand `smore eat` is applied on output of `smore roast`. Here, the output folder of `smore roast` is the input to `smore eat` which will take into account all files in the folder with having specific names as given by `smore roast`. This last part of the program will then summarize all the information such that evolutionary events can be counted and numbers added to the tree. The help page for `smore eat` is `smore eat --help`.

parameter	obl/opt	description
`--out|-o OUTPATH`	obl	directory where to write the output files. If the folder does not exist, it will be created.
`--prep PATH`	obl	directory where `smore prep` output files are located.
`--mix PATH`	obl	directory where `smore mix` output files are located.
`--roast PATH`	obl	directory where `smore roast` output files are located.
`--newick NEWICK_FILE`	obl	phylogenetic tree in newick format including all species contained in the analysis. For automatic conversion of species identifiers, use parameter –id.
`--id ID_FILE`	opt	in case species identifier in the tree and multiZ files do not match, this file can be used to automatically convert species identifier. Format of the table: current_name_in_tree (tab) name_to_translate_to
`--nomiss`	opt	using this option, the pipeline will omit the check for missing data. Hence, deletions in the tree won't distinguish between a deletion because of missing data or a 'real' deletion. Only useful when running on a large data set.

`--tool|-t TOOLPATH`	opt	directory where *SMORE* is located. This parameter can usually be omitted.
`--python PYTHON_PATH`	opt	Path to `python` installation if the one specified in the path environment is not used.
`--perl PERL_PATH`	opt	Path to `perl` installation if the one specified in the path environment is not used.

5.7 Summary of Subcommands

table with subcommands and short explanation of parameters

6 Example

Take example of tRNAs 6 species with tRNAscan

example of prep-mix-roast-eat

7 Theory

8 Help Pages

9 Contact

References

[1] Velandia-Huerto, C.A., Berkemer, S.J., Hoffmann, A., Retzlaff, N., Romero Marroquín, L.C., Hernández Rosales, M., Stadler, P.F., Bermúdez-Santana, C.I.: Orthologs, turnover, and remolding of tRNAs in primates and fruit flies. BMC Genomics **17**, 617 (2016). doi:10.1186/s12864-016-2927-4

[2] Berkemer, S.J., Hoffmann, A., Murray, C.R., Stadler, P.F.: SMORE: Synteny Modulator Of Repetitive Elements. MDPI Life (2017 submitted)

[3] Nawrocki, E.P., Eddy, S.R.: Infernal 1.1: 100-fold faster rna homology searches. Bioinformatics **29**, 2933–2935 (2013)

[4] Lowe, T.M., Eddy, S.R.: tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. **25**, 955–964 (1997). doi:10.1093/nar/25.5.0955

[5] Letunic, I., Bork, P.: Interactive tree of life (itol) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic acids research **44**(W1), 242–245 (2016)