SMART pipeline Documentation

Release 1.0

M. Sejourne, A. Teissander, G. Boldina, M. Dutertre, N. Servant

April 22, 2015

CONTENTS

Contents:

CHAPTER

ONE

SMART QUICK START GUIDE

See NEWS for information about changes in this and previous versions

See LOGBOOK for details about the HiC-Pro developmens

1.1 What is the SMART pipeline ?

SMART (3'-Seq Mapping Annotation and Regulation Tool) was set up to process 3' sequencing data from aligned sequencing reads. It includes the peak detection, filtering, and annotation. If you use it, please cite : *Boldina et al. 2015*

1.2 How to install it ?

The following dependancies are required : * * R with the *RColorBrewer*, *ggplot2*, *rtracklayer*, *DESeq2*, *magrittr*, *dplyr*, *qplots* and *genomicRanges* packages * Samtools (>0.1.18) * BEDTools (>2.17.0)

To install the SMART pipeline, simply extract the archive and set up the configuration file with the paths to dependencies.

```
tar -zxvf smart_1.0.0.tar.gz
cd smart_1.0.0
```

1.3 Annotation Files

The pipeline is using a couple of annotation files with gene annotation and last exons information. These files are based on UCSC Refseq gene. In order to generate all required annotation files, please set the ANNOT_DIR, ORG and UCSC_EXPORT in the configuration file

```
BUILD_ANNOT=1
ORG=mm9
UCSC_EXPORT=refseq_export_mm9.csv
```

SMART will start by creating the annotation files in the forder ANNOT_DIR/ORG/ The annotation only have to be generated once. Mouse annotation are provided with the pipeline.

1.4 How to use it ?

SMART can be used both for a single sample or for a list of samples. In the case of sample list, peaks detected in all samples are merged before annotation. In oder to use the pipeline, please set up the configuration according to your analysis, and run the following command :

```
/bin/smart -c CONFIG [-i INPUT_BAM] [-l INPUT_LIST] -o OUTPUT_DIR
```

CHAPTER

TWO

SMART PIPELINE MANUAL

2.1 What is the SMART pipeline ?

SMART was set up to process PolyA sequencing data from aligned sequencing reads. It includes the peak detection, filtering, and annotation. If you use it, please cite : *Boldina et al. 2015*

2.2 How to install it ?

The following dependancies are required : * * R with the *RColorBrewer*, *ggplot2*, *rtracklayer*, and *genomicRanges* packages * Samtools (>0.1.18) * BEDTools (>2.17.0)

To install the SMART pipeline, simply extract the archive.

```
tar -zxvf smart_1.0.0.tar.gz
cd smart_1.0.0
```

Then, edit the *config.txt* file and manually defined the paths to dependencies.

SYSTEM CONFIGURATION	
SAMTOOLS_PATH	Full path to the samtools installation directory (>0.1.18)
BEDTOOLS_PATH	Full path to the BEDTools installation directory (>2.17.0)
R_PATH	Full path to the R installation directory
PYTHON_PATH	Full path to the python installation directory
AWK_PATH	Full path to the awk utility

2.3 Annotation Files

The pipeline is using a couple of annotation files with gene annotation and last exons information. These files are based on UCSC Refseq gene. This file can be downloaded from the UCSC TableBrowser website and should look like this :

```
#bin  name    chrom   strand  txStart txEnd   cdsStart
cdsEnd        exonCount       exonStarts
exonEnds      score   name2   cdsStartStat    cdsEndStat      exonFrames
0     NM_001195025    chr1    +       134212701       134230065       134212806
      134228958       8       134212701,134221529,134222782,134224273,134224707,
      134226534,134227135,134227897,  134213049,134221650,134222806,134224425,
      134224773,134226654,134227268,134230065,        0       Nuak2   cmpl
      cmpl    0,0,1,1,0,0,0,1,
0     NM_028778       chr1    +       134212701       134230065       134212806
```

```
      134228958       7       134212701,134221529,134224273,134224707,134226534,
      134227135,134227897,    134213049,134221650,134224425,134224773,134226654,
      134227268,134230065,    0       Nuak2   cmpl    cmpl    0,0,1,0,0,0,1,
```

In order to generate all required annotation files, please set the ANNOT_DIR, ORG and UCSC_EXPORT in the configuration file

```
BUILD_ANNOT=1
ORG=mm9
UCSC_EXPORT=refseq_export_mm9.csv
```

The pipeline will start by creating the annotation files in the forder ANNOT_DIR/ORG/ The annotation only have to be generated once. Mouse annotation are provided with the pipeline.

In addition, SMART will look for known polyA signal in the detected peaks and flanking regions. This list of motif is defined in the *polyA_signal.csv* file and can be edited.

```
AATAAA
ATTAAA
AGTAAA
TATAAA
CATAAA
AAGAAA
GATAAA
AATATA
AATGAA
TTTAAA
AATACA
AAAAAG
ACTAAA
AATAGA
```

2.4 How to use it ?

SMART can be used both for a single sample or for a list of samples. In the case of sample list, peaks detected in all samples are merged before annotation. In oder to use the pipeline, please set up the configuration according to your analysis, and run the following command :

```
/bin/smart -c CONFIG [-i INPUT_BAM] [-l INPUT_LIST] -o OUTPUT_DIR
```

2.5 How to use it ?

1. Copy and edit the configuration file *'config.txt'* in your local folder. The '[' options are optional.

GENE ANNOTATIONS	
BUILD_ANNOT	0/1 - Run the annotation builder
ORG	Organism
UCSC_EXPORT	UCSC reference to build the annotation (i.e. refseq_export_mm9.csv)
POLYA_MOTIF	List of polyA annotation signal (i.e annotation/polyA_signal.csv)

PEAK DETECTION	
MIN_MAPQ	Minimum reads mapping quality (default: *20*)
MAX_DIST_MERGE	Maximum distance between reads to be merged as a peak (default: *170*)
MIN_NB_READS_PER_PEAK	Minimum number of reads per peads (default: *5*

PEAK FILTERING	
NB_STRETCH_POLYA	Window length to look for polyA stretch (default: *9*)
MISM	Number of non-A base allowed in the NB_STRETCH_POLYA window (default: *1*)
NB_STRETCH_CONSECUTIVE	Minimum size of A stretch (default: *6*)
WINSIZE_DOWN	Window size downstream the peaks (default: *150*)
WINSIZE_UP	Window size upstream the peaks (default: *50*)
KEEP_LE_PEAKS	Always keep peaks in gene's last exon (0/1, default: *1*)

ANNOTATION	
MIN_LE_OV	Minimum overlap to consider a peak as overlapping with a last exon (default: *1*)
MIN_INTRON_OV	Minimum overlap to consider a peak as overlapping with a intron (default: *3*)
MIN_ANNOT_OV	Minimum overlap for peaks annotation (default: *1*)

SAMPLES COMPARISON	
COMBINE_SAMPLE	Samples to merge before comparison. Should be under bracket, with comma separated (i.e [1,2,3,4][5,6,7,8] ...)
COMPARE_SAMPLE	Define group to compare. Must be defined as COMBINE_SAMPLE with group 0 vs group 1 (i.e [0,0,1,1][0,0,1,1] ...)
MIN_COUNT_PER_COND	Minimum sum of counts per condition. (Default: *10*)

2. Edit the sample list files in case of multiple samples. This file must be tab delimited with *sample_number* t *file* t *sample_id*. Note that the *sample_number* must correspond to the COMBINE_SAMPLE variable from the configuration file. These samples will be combined to define a common set of peaks which can further be used for differential analysis. The *sample_id* are used for the differential analysis only. Here is an sample list file example :

:: 1 /data/sample1.bam COND1 2 /data/sample1.bam COND1 3 /data/sample1.bam COND2 ...

3. Run the pipeline

- For one file

```
/bin/smart -c CONFIG -i INPUT_BAM -o OUTPUT_DIR
```

- For a list of file

```
/bin/smart -c CONFIG -l INPUT_LIST -o OUTPUT_DIR
```

2.6 How does SMART work ?

TODO