ncPRO-seq

Annotation and profiling of ncRNAs in sRNA-seq data

User Guide

ncPRO-seq 1.6.5

Chong-Jian Chen & Nicolas Servant

Last updated : February 2, 2016

Contents

1 Overview

1.1 Before starting

Over recent years, deep sequencing technology has become a powerful approach for deeply investigating small non-coding RNA (ncRNA) populations, i.e. small RNA-seq. It is now established that an increasing number of novel small ncRNA families distinct from microRNAs are generated over kingdoms from different coding/non-coding regions via various biogenesis pathways and might involve a great spectrum of biological processes. For example, two other major classes of endogenous small RNAs, Piwi-interacting RNAs (piRNAs) and endogenous small interfering RNAs (endo-siRNAs), have been identified and widely investigated in mammals [9]. Moreover, in other organisms like plants more classes of small ncRNA have been described indicating that a wide range of small ncRNAs exist [5].

However, most of the existing tools devoted to sRNA-seq analysis, are only based on miRNAs annotation and quantification, or can only be applied to one organism. Here we present a comprehensive and flexible ncRNA analysis pipeline, **ncPRO-seq 1.6.5** (**N**on-**C**oding RNA **PRO**filing in sRNA-**seq**) (http://ncpro.curie.fr/), which is able to interrogate and perform detailed analysis on small RNAs derived from annotated non-coding regions in miRBase [12], Rfam [7] and repeatMasker [18], and regions defined by users. The ncPRO-seq 1.6.5 pipeline also has a module to identify regions significantly enriched with short reads that can not be classified as known ncRNA families. The ncPRO-seq pipeline supports input read sequences in fastq, fasta and color space format, as well as alignment results in BAM format, meaning that small RNA raw data from the 3 current major platforms (Roche-454, Illumina-Solexa and Life technologies-SOLiD) could be analyzed with this pipeline. Finally, the ncPRO-seq pipeline can be used to analyze data based on genome from metazoan to plants.

1.2 About us

ncPRO-seq 1.6.5 is developped in the context of a collaborative project involving the following partners :

- Institut Curie - Platforme Bioinformatique (France, Paris)
- Génétique et biologie du développement (France, Paris) - Institut Curie - UMR 3215 CNRS - U934 Inserm
- Arabidopsis Epigenetics and Epigenomics group (France, Paris) - CNRS UMR8197 - INSERM U1024 - Institut de Biologie de l'Ecole Normale Sup?rieure
- Institut de Biologie Mol?culaire des Plantes du CNRS - UPR2357 (France, Strasbourg)
- Department of Biology - Swiss Federal Institute of Technology Zurich (Suisse, Zurich)

If you use the ncPRO-seq tool for your analyses, please cite the following paper :
Chen C., Servant N., Toedling J., Sarazin A., Marchais A., Duvernois-Berthet E., Cognat V., Colot

V., Voinnet V., Heard E., Ciaudo C. and Barillot E. *ncPRO-seq: a tool for annotation and profiling of ncRNAs in sRNA-seq data.* submitted.

2 Requirements and Installation

2.1 Hardware

Computer with at least 4GB of primary memory is recommended. Although the ncPRO-seq 1.6.5 pipeline is fast in a simple desktop computer or laptop, launching the pipeline in a computer cluster would achieve much better performance by turning on the multithreading option.

2.2 Operating system

In the current version, the ncPRO-seq 1.6.5 pipeline can only be installed in a Linux/UNIX-like operating system (Linux/UNIX or Mac OS X).

2.3 Required softwares

The ncPRO-seq 1.6.5 pipeline requires the following additional softwares. Please refer to section 8 for more details :

- The Bowtie Aligner (<v2.0) [13] to align the reads from smallRNA-seq data.
- The R (>=v3.2.3) [17].
- The R and BioConductor [8] packages: seqLogo [3], girafe [20], RColorBrewer [15], ggplot2 [22], reshape2 [21] and gridExtra [2].
- The BEDTools suite (>=v2.15.0) [16], and the bamMapCount program to address common genomic tasks such as finding feature overlaps and grouping same features.
- The SAMTools suite [14] to handle SAM and BAM format.
- The Convert/ImageMagick utilies to format files and images.
- The Python (>=v2.6.6).
- The Python package: ReportLab (>=v2.7).
- In order to use the end-user interface, a local server with php such as apache is needed.
- For Mac OS X user, make sure that you have Xcode and command line tools for Xcode installed in your system.

2.4 Installation of ncPRO-seq 1.6.5

1. Make sure that R has been successfully installed in your system. If you want to use the interface version of ncPRO-seq 1.6.5, a local server with php is also required before the ncPRO-seq 1.6.5 installation. Please refer to section 8 for more details of these two softwares.

2. Download and extract pre-built bowtie genome index files from bowtie website or build your own genome indexes by following the instructions in bowtie website. The path of bowtie indexes will be demanded during the following ncPRO-seq 1.6.5 installation.

3. Download the source of the latest version of ncPRO-seq 1.6.5 and annotation files of different species from sourceforge (http://sourceforge.net/projects/ncproseq/files/). For example, we choose source code **ncPRO-seq.v1.6.1.tar.gz** from src folder and annotation file **mouse_mm9.tar.gz** from annotation folder in the ncPRO-seq 1.6.5 download page in sourceforge. Then go to the directory where you save downloaded files in your shell prompt, do the following steps to extract files and to put annotation files into the source folder.

   ```
   > tar -zxvf mouse_mm9.tar.gz
   > tar -zxvf ncPRO-seq.v1.6.1.tar.gz
   > mv mm9 ncPRO-seq.v1.6.1/annotation/
   ```

4. Use the following command lines to semi-automatically install the command line version and/or the interface version of ncPRO-seq 1.6.5, as well as required softwares except R and the local server as mentioned in section 2.3. During the installation, users need follow the guides to choose or input options. Make sure that you have access to internet at the shell prompt under current user's privilege.

   Common users without super-user privileges can only install the command-line version of ncPRO-seq 1.6.5, as configurations of local web server need super-user permissions. The installation command is the following:

   ```
   > cd ncPRO-seq.v1.6.1
   > make install
   ```

   For the super-user or root user, the following commands can be used to install both command-line and local web server version of ncPRO-seq 1.6.5:

   ```
   > cd ncPRO-seq.v1.6.1
   > sudo make install
   ```

   or

   ```
   > su
   > make install
   ```

5. The user just has to run the following script to test the successful installation of ncPRO-seq 1.6.5. Note that you need change **ncpro-install-dir** to the path where you choose to install ncPRO-seq 1.6.5 during the installation.

```
> ./check_install.sh -p ncpro-install-dir
```

After the installation, the `config-system.txt` file is generated and can be found in the folder where you installed ncPRO-seq 1.6.5. This file lists all paths of environment and softwares required by the ncPRO-seq 1.6.5. Users are not recommened to make any changes in this file, unless you are an expert or are aware of some errors.

Options	Description	Example
INSTALL_WWW	0/1. If 1, the web interface of ncPRO-seq 1.6.5 is installed in **WWW_DIR** and **CGI_DIR**	1
APPLI_DIR	Absolute path of installation folder	/home/ncPROseq/
WWW_DIR	HTML installation folder. Have to be under the web working directory	/var/www/ncPROseq/
CGI_DIR	CGI installation folder. Have to be under the CGI working directory	/usr/lib/cgi-bin/ncPROseq/
WWW_RES	Results path for local web version	/home/ncPROseq/www_results
PBS_MODE	0/1. If 1, the job are send to the PBS job scheduler using the *qsub* command	0
PBS_OPT	Options to give to *qsub* at job submission	-m ae -M bioinfo.ncproseq@curie.fr -j oe -l nodes=1:ppn=6,mem=20gb
PBS_PATH	Absolute path to PBS (qsub) command	/usr/bin
R_PATH	Absolute path to R binary folder	/usr/local/R/R/bin/
AWK_PATH	Absolute path to awk executable folder	/usr/bin/
BOWTIE_PATH	Absolute path to bowtie executable folder	/usr/local/bowtie-0.12.7/
BOWTIE_INDEXES_PATH	Installation folder of Bowtie indexes	/home/Apps/Bowtie_indexes/

BEDTOOLS_PATH	Installation folder of BedTools binaries	/home/Apps/BEDTools-Version-2.15.0/bin/
SAMTOOLS_PATH	Installation folder of samtools binary	/home/Apps/samtools/
CONVERT_PATH	Utilities used to create thumbnail for html report	/usr/bin/
BAM_MAPCT_PATH	Installation folder of the bamMapCount program. Distributed with the BEDTools-version10.0	/home/Apps/BEDTools-Version-2.10.0/bin/
PERL_PATH	Perl installation folder	/usr/local/bin
PYTHON_PATH	Python installation folder	/usr/local/bin
MAILX_PATH	Optional - mailx excutable path to send email at the end of the analysis	/usr/bin/

Table 1: Paths in one example of config-system.txt

3 How to use it ?

3.1 Configuration file

ncPRO-seq 1.6.5 is a flexible pipeline which allows users to specify different options at each analysis stage, from raw reads processing to ways to generate results. In the web interface version, all options can be easily chosen through the web (see 3.2). In command line version (see 3.3), users can manually edit the `config-ncrna.txt` file to define options according to the descriptions of options below (Table 6) and also inside the file. In the `config-ncrna.txt` file, you may find more options than that in the web page, especially for the Bowtie mapping step, but we do not suggest you to make any changes in these extra options unless you are an expert.

Options	Description
N_CPU	Number of CPU used by Bowtie to do mapping
LOGFILE	File that lists actions that have occurred during the analysis
PROJECT_NAME	Project name for the report.

FASTQ_FORMAT	The quality score format of the fastq reads. Three formats are supported: **phred33** (Sanger, Solexa version 1.8 or later), **solexa** (Solexa prior to version 1.3), **solexa1.3** (Solexa version 1.3 to 1.7)
BOWTIE_GENOME_REFERENCE	Basename of the Bowtie index genome reference file (base space). See the Bowtie manual for additional informations
BOWTIE_GENOME_REFERENCE_CS	Basename of the Bowtie index genome reference file (color space). See the Bowtie manual for additional informations
BOWTIE_GENOME_OPTIONS_FQ	Options for Bowtie to map base space reads in fastq format (Solexa)
BOWTIE_GENOME_OPTIONS_FA	Options for Bowtie to map base space reads in fasta format (454)
BOWTIE_GENOME_OPTIONS_CS	Options for Bowtie to map color space reads (SOLiD)
GROUP_READ	Group reads based on their sequence for raw reads before mapping or read alignments in bam file depending on the input format. **1**: Yes; **0**: No; **2**: for the online version where the input files have already been grouped using our provided scripts
FILTER_ZERO_COUNTS	mRNAs that do have in at least one sample a count above zero in count files. **1**: Yes; **0**: No, all miRNAs in count files.
ORGANISM	Name of the reference organism. Must be the same as the organism available in the annotation folder (i.e. mm9, hg19, ...)
MATURE_MIRNA	Annotation against miRNAs from miRBase. Both miRNA with and without an extended item are acceptable (see 5.4.2)
PRECURSOR_MIRNA	Annotation against pre-miRNAs from miRBase. Both miRNA with and without an extended item are acceptable (see 5.4.2)
NCRNA_RFAM	List of the RFAM ncRNA(s) to focus on (comma separator) - no extension parameter
NCRNA_RFAM_EX	List of the RFAM ncRNA(s) to focus on (comma separator) - extension parameter (see 5.4.2)
NCRNA_RMSK	List of the repetitive elements to focus on (comma separator) - no extension parameter

NCRNA_RMSK_EX	List of the repetitive elements to focus on (comma separator) - extension parameter (see 5.4.2)
TRNA_UCSC	Mapping against tRNA sequences. Both tRNA with and without an extended item are acceptable (see 5.4.2)
OTHER_NCRNA_GFF	List of custom gff files to intersect with the mapped reads
LOGO_DIRECTION	Align the sequence on the 5' or 3' end **[5/3]**
IC_SCALE	Use the information content scale for Logo outputs. **1**: Yes; **0**: No
GENOME_TRACK_OPTIONS	Options to select reads mapped in the genome to generate track file. Four options should be provided to filter reads, and separated by comma. **min_len=N** : the minimum length (N) of read; **max_len=N**: the maximum length (N) of read; **min_copy=N** : the minimum number (N) of matches in the genome; **max_copy=N**: the maximum number (N) of matches in the genome. To have more than one type of track, different sets of options should be separated by pipe (|)
SIG_READ_OPTIONS	Options to select mapped reads for enrichment analysis (see 5.5). Please refer to the format of **GENOME_TRACK_OPTIONS**
SIG_WIN_SIZE	The window size used to scan the genome (e.g. 10000) (see 5.5)
SIG_STEP_SIZE	The step size (e.g. 50000) (see 5.5)
EXCLUDE_ANN_GFF	List of annotation files (gff3). Only reads which are not mapped in these annotated regions are kept for enrichment analysis (see 5.5)
FIT_MODEL	The model used to fit window-based read distribution. Three models can be chosen: **NB.ML**, **NB.012**, and **Poisson** (see 5.5)
PVAL_CUTOFF	The cut-off used to get regions significantly enriched with reads

Table 2: Options from the configuration file

3.2 Web interface version

The ncPRO-seq 1.6.5 pipeline provides a local web interface to run the analysis. Most of the options describe in the section 3.1 can be defined using the interface. To open the ncPRO-seq 1.6.5 interface, just enter the web url (as defined in `WWW_DIR` in the configuration system file), followed by the `load.php` file. For instance, open http://localhost/ncPROseq/load.php in your favourite navigator.
This web implementation uses a Perl CGI scripting program, and required perl to be installed in /usr/bin/perl. This path is the default path under Unix system. If not available, please ask your administrator to create a link to your own perl environment.

3.2.1 Input data files

The local input files can be selected as shown in Figure 1. As previously described, the .fastq, .csfasta, and .fasta raw sequencing data can be selected, as well as .bam files for reads already aligned on the genome. These different file types can be mixed in a single analysis.
If a folder is selected, all the files within the folder will be analysed. In order to avoid security issues (i.e. browsing the entire file server), **the user has to enter the full path of its data**. As any standard web application, ncPRO-seq 1.6.5 is run by the http user. Accordingly, **it is your responsabilities to check if the http user can read your data**.

3.2.2 Alignment on a reference genome

The Figure 2 presents the different parameters to set for the short reads mapping using the Bowtie alignment tool.
First, the user has to select if the reads has to be grouped before mapping (see section 5.2.1). We highly recommended to use this option, as it reduces a lot the size of the input reads (and output files), and have only very few impact (if any) on the results. All the reads corresponding to the same sequence are merged, and the related quality values are averaged.
The following Bowtie options can then be available. For details explanation, please see the Bowtie manual page.

- **Bowtie pre-built index.** The name of the index file to be searched in the Bowtie index folder (as specified in the system configuration file during installation).
- **Mapping options.** Using the interface, all input file are aligned with the options '-a –best –strata -y' to always select the best hits. Then, the number of mismatches (up to 3), the maximum number of reportable alignments or the fastq format have to be choose according to the input data type, and the goal of the analysis.
- **Bowtie multithreading option.** Number of CPUs used by Bowtie to perform the alignment.

Finally, the user can also ask for a quality mapping report. In this case, the following outputs are generated :

- Reads length distribution.

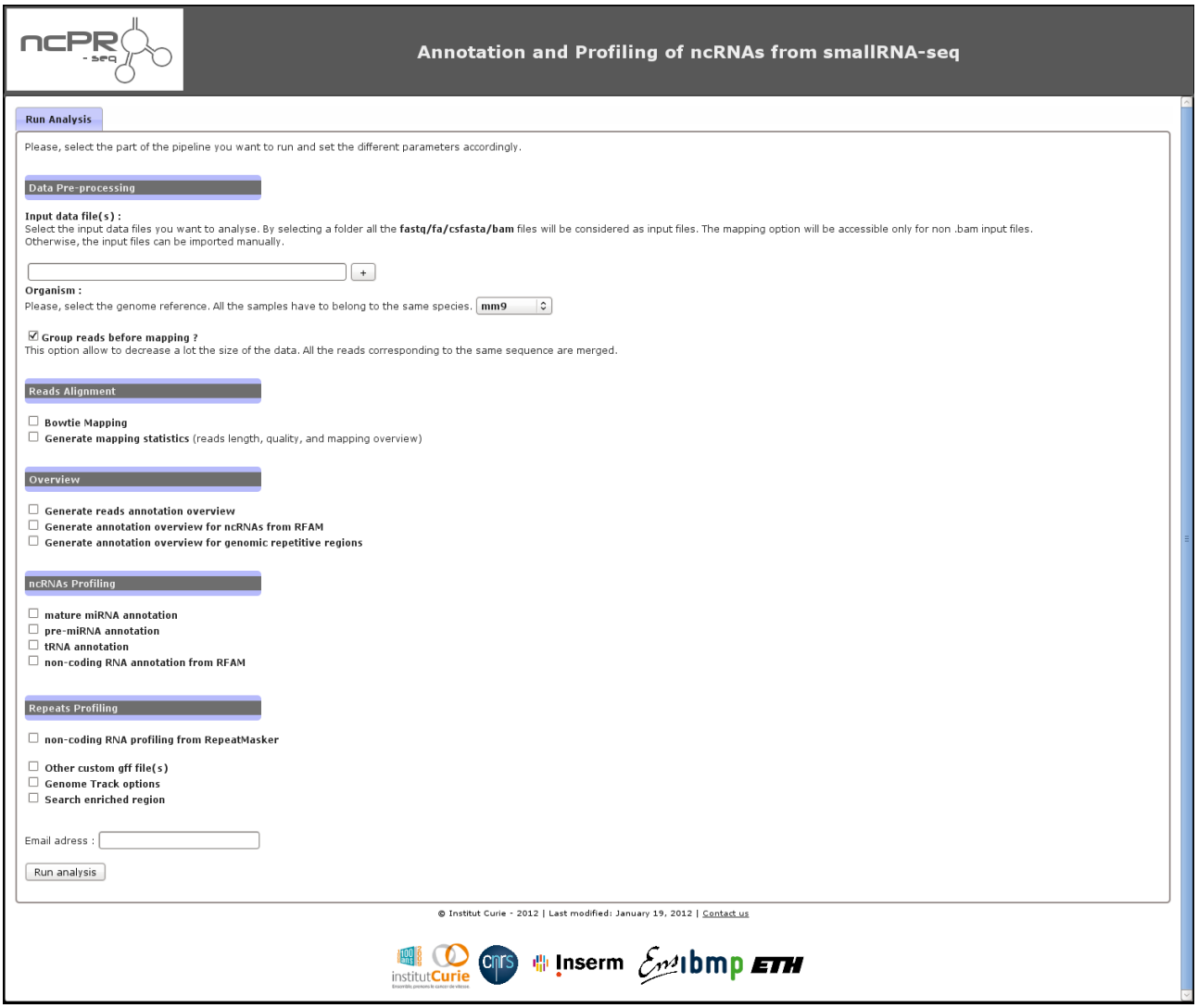

Figure 1: ncPRO-seq 1.6.5 web interface: load input files

- Sequence length distribution (only the distinct reads are used).
- Number of reads aligned on the genome reference.

3.2.3 Annotation Overview

The overview section aims in annotating the reads and classifying them into large families (see the section 5.4.5).
The reads annotation family is the most general view, and annotates the reads based on the following annotations : coding genes, ncRNAs from Rfam, smallRNAs from repeated regions, rRNA, and precursor miRNA from miRBase. Please, note that ncPRO-seq 1.6.5 uses the miRBase annotation for miRNAs, and that accordingly, the miRNAs from Rfam are not used.

Figure 2: ncPRO-seq 1.6.5 web interface: set alignment parameters

Then, an overview of the non-coding RNAs annotation and of the repetitive genomic regions are provided.

Figure 3: ncPRO-seq 1.6.5 web interface: reads annotation

3.2.4 Profiling of non-coding RNAs

Then, the profiling of the non-coding RNAs (section 5.4.6) can be performed by selecting the RNAs classes to focus on and the associated extension parameters as presented in section 5.4.2 (Figure 4). Regarding the ncRNAs from the RFAM database, several targets can be specified. The profiling analysis will be separately done for each input keyword.

3.2.5 Profiling of repetetive regions

The same profiling options are available for the annotation of the repetitive genomic regions (Figure 5). Here, ncPRO-seq 1.6.5 offers the possibility to select only full length repeated elements. The full length elements are detected using the position of each element on the consensus sequence (data from UCSC). Only the elements matching the exact length of the consensus are considered as full length.

Figure 4: ncPRO-seq 1.6.5 web interface: annotation of ncRNAs

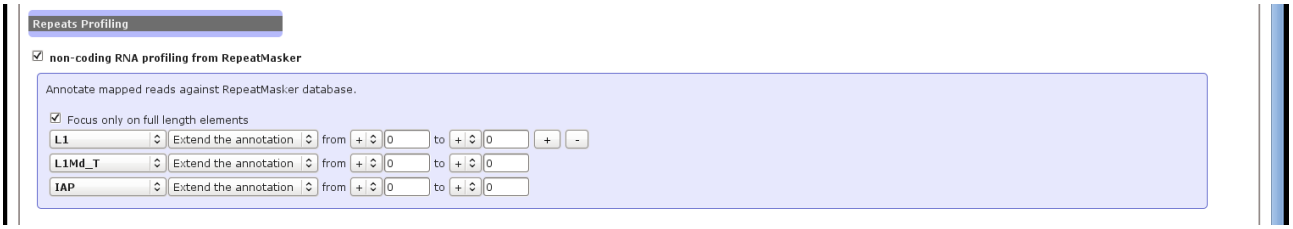

Figure 5: ncPRO-seq 1.6.5 web interface: annotation of repetitive regions

3.2.6 User Custom gff files

For each annotation family including different repeat/Rfam families and other custom annotations, ncPRO-seq 1.6.5 creates a table file showing read coverages of each single family member in all sequencing libraries. If some annotations are not available, or if the user want to search for reads annotated in other genomcis regions, the profiling can also be done on custom gff files. As for the input files, the full path of the gff files has to be set. Several custom gff files can be specified.

3.2.7 Genome tracks visualisation

For each sequencing library, for each annotation family, ncPRO-seq 1.6.5 generates two compressed track files in bedgraph and bed formats to describe read mapping in sense and antisense direction of annotation items respectively. These track files can be then, easily upload in a visualisation tool as the UCSC Genome Browser [6].
Moreover, the interface allows users to select different subsets of reads mapped in the genome to generate BedGraph formatted track files. The minimum and maximum size of the reads, as well as the minimum and maximum number of locations can be used to filter the aligned reads (Figure 6).

Figure 6: ncPRO-seq 1.6.5 web interface : genome tracks options

3.2.8 Search enriched regions

The enrichment analysis can be launch from the interface, by specifying the different options as explained in the section 5.5. As for the Genome tracks options, the user can choose to focus on a subset of reads. The minimum and maximum size of the reads, as well as the minimum and maximum number of locations can be used to select these reads. Then only the reads not annotated on the selected items (coding genes, miRNAs, ncRNAs) will be used for the analysis (Figure 7).

Figure 7: ncPRO-seq 1.6.5 web interface : search enriched regions

3.2.9 Run the analysis

Finally, after setting the different parameters, the ncPRO-seq 1.6.5 pipeline can be executed. A random url will be generated (Figure 8) to visualize the output report (see section 4). If specified, an email will be send to the user at the end of the analysis.
Please, note that in the current version the report will be updated only **at the end** of the analysis.

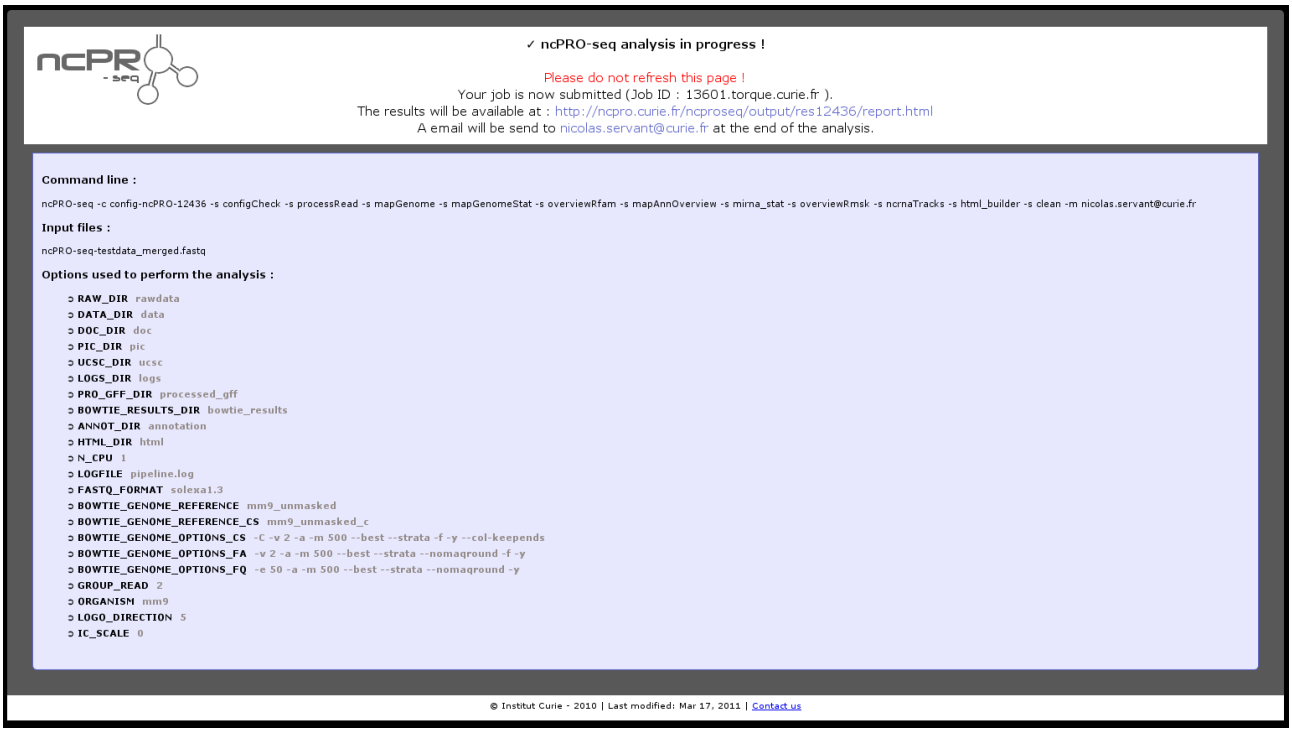

Figure 8: ncPRO-seq 1.6.5 web interface: run the analysis

3.3 Command-line version

The ncPRO-seq 1.6.5 pipeline will generate a lot of output files. Thus before starting, it is highly recommended to deploy the ncPRO-seq 1.6.5 output architecture.
However, this step is **optional**. If you don't want to use this output architecture, please change the output paths in the configuration files. To deploy the output architecture, use the following command:

```
$ MY_INSTALLATION_DIR/bin/ncPRO-deploy -o MY_OUTPUT_DIR
```

The input (fastq, bam, csfasta or fasta) files must be filed in the rawdata folder. Finally, after setting the different parameters in the configuration file, run the ncPRO-seq pipeline as follow :

```
$ cd MY_OUTPUT_DIR
$ MY_INSTALLATION_DIR/bin/ncPRO-seq -c config-ncrna.txt
```

The ncPRO-seq 1.6.5 pipeline is modular and sequential. The user can specify the analysis steps to run. For instance, the following command line will just perform the quality control, the reads grouping, and the alignment on the reference genome.

```
$ MY_INSTALLATION_DIR/bin/ncPRO-seq -c config-ncrna.txt -s processRead
-s mapGenome -s mapGenomeStat
```

The following analysis steps are available:

ncPRO-seq 1.6.5 analysis step (-s option)	Description
processRead	calculate read length distribution, median quality score for each postion, and group reads
processBam	processed and group reads from bam files
mapGenome	run Bowtie for genome mapping
mapGenomeStat	compute number of mapped reads and unmapped reads in the genome
mapAnnOverview	compute overview of reads annotation
overviewRmsk	compute read coverage for each repeats family
overviewRfam	compute read coverage for each ncRNA family
generateNcgff	create gff file for special ncrna family
ncrnaProcess	ncRNA family analyses, including read coverage, read length distribution, read coverage in subfamilies, and sequence logo
genomeTracks	genreate genome coverage ucsc track
ncrnaTracks	genreate ucsc tracks for ncrna visualization
sigRegion	detect significantly enriched regions
html_builder	build the html report file
pdf_builder	build the pdf report file

Table 3: Description of ncPRO-seq 1.6.5 '-s' options

4 How to browse the results ?

Users just need to open "report.html" automatically created by ncPRO-seq 1.6.5 in a web browser to easily view figures and tables which are originally stored in `pic` (for pictures) and `doc` (for table files) folder respectively, and to access track files in UCSC folder (Figure 9). The report file is composed of 7 types of tab. Briefly, each tab presents the pictures (or tables) generated by the pipeline. Each picture, and table can be visualized in high resolution, and/or download for further analysis.

Home. The main page of the report list all samples and options used to perform the analysis, and the version of pipeline, used softwares and annotation files.

Quality Control. All the quality controls performed on the raw input data are presented in this tab. The mean quality score, base, GC, and the insert length distribution are available.

Data Mapping. As for the quality control, all the pictures regarding the alignment of reads on the reference genome are presented here. The reads length distribution, and the mapping proportion are available. These first controls give a good idea of the overall quality of the input libraries.

ncRNAs Overview. The annotation overview of the different ncRNAs family is separated in pre-miRNAs, rRNA, repeat, rfam, protein coding gene and unknown. For the RFAM/repeats annotation, the ncPRO-seq 1.6.5 pipeline count the number of abundant reads in each family and provides the relative proportion.

ncRNAs Profiling. Then, for each ncRNAs specified in the configuration file (see section 5.4.2), and each samples, ncPRO-seq 1.6.5 generate the coverage profile and the logo sequences. As for most of the analysis, both results at the reads or the sequence level are available. Regarding the logo sequences, both view (all sequences, or major sequence) are available (see section 5.4.7).

Table view. All the table files generated by the ncPRO-seq 1.6.5 pipeline can be browsed, visualized, and downloaded. The files are organized from general overview, to ncRNAs profiling information.

Genome Tracks. The bedgraph or bed files generated by the pipeline can be downloaded and loaded in a standard visualisation tool, such as the UCSC browser [6]

Logs. The log file of the ncPRO-seq 1.6.5 process is printed here. Check this file in case of error.

Help. This manual is available trough the report interface.

Figure 9: ncPRO-seq 1.6.5 HTML analysis report

5 How does-it work ?

5.1 Structure of the pipeline

The workflow of the ncPRO-seq 1.6.5 pipeline consists of five main components: **input pre-processing**, **read mapping**, **read annotation**, **annotation analyses**, and **enrichment analyses**, as illustrated in Figure 10 with different background color. We will explain the tasks, key technical points and results of each step in the following subsections.

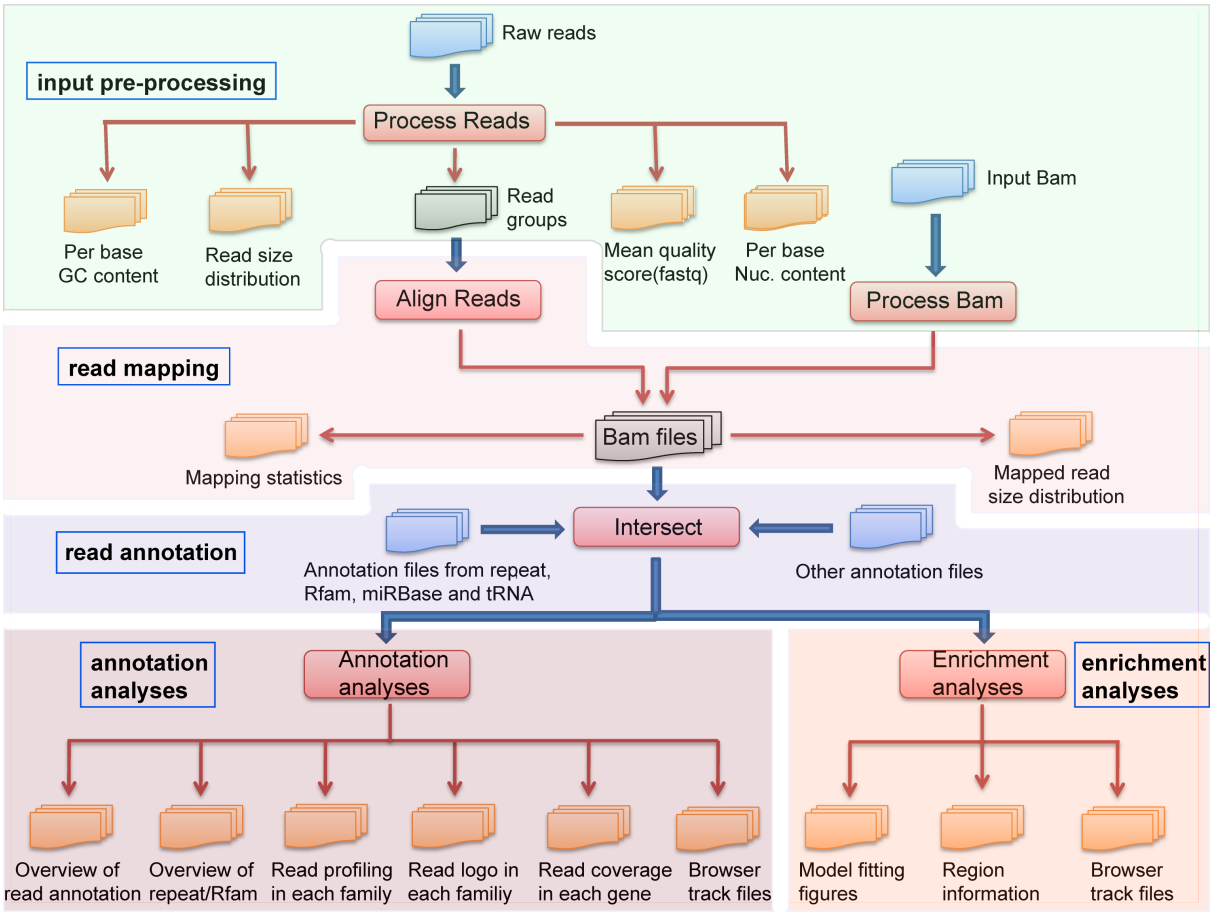

Figure 10: The whole workflow of ncPRO-seq 1.6.5

5.2 Input pre-processing

Both read sequences and alignment results are acceptable as input in ncPRO-seq 1.6.5. Input files are pre-processed before being sent to the next step.

5.2.1 Read pre-processing

ncPRO-seq 1.6.5 supports input read sequences in three formats: **fastq** from Solexa, **csfasta** from SOLiD and **fasta** from 454. In this step, ncPRO-seq 1.6.5 generates several figures for each sequencing library to describe the basic properties of sequencing reads, such as length distribution of distinct reads and abundant reads, positional base and GC content, and positional mean quality score if the read format is fastq, all of which are useful to access the basic quality of sequencing reads. Note that **distinct reads** are read groups that only count once for reads with the same sequence, i.e. ignoring the read abundance, whereas **abundant reads** are all sequenced reads.

There is an useful option in ncPRO-seq 1.6.5, called `GROUP_READ` (see 3.1), which controls the read clustering process. If it is set to 1, reads with identical sequence will be clustered into non-redundant read groups which are then specified with unique group id and read count. For reads in fastq format, the positional quality score of a read group is the mean positional quality score of all reads that are clustered in this group. In the following analyses, read groups, which has a significant decrease of read items comparing to the original read data, will be processed instead as input data. We recommend users to use this option especially if the sequencing libraries are extremely big, which will significantly reduce the CPU time for all analyses and the disk space as well to store intermediate results. Furthermore, another advantage of using this option is that you will get additional read profiles computed based on read groups (i.e. distinct reads) as shown in 5.4.6

5.2.2 Alignment pre-processing

ncPRO-seq 1.6.5 can also take read alignment files in BAM format as input. BAM is the compressed binary version of the SAM format, please check here for more information. The `GROUP_READ` option also works for BAM file input. The clustering process is based on read sequences in alignments using the similar method as described in 5.2.1. After this step, a new bam file will be generated by replacing original read information with read group and removing redundant alignments in original bam file.

5.3 Reads mapping

In ncPRO-seq 1.6.5, Bowtie (<v2) [13] is used to align reads to the reference genome. Different bowtie options can be specified to map different formats of reads: `BOWTIE_GENOME_OPTONS_FQ` for fastq reads, `BOWTIE_GENOME_OPTIONS_FA` for fasta reads and `BOWTIE_GENOME_OPTIONS_CS` for csfasta reads. Bowtie outputs alignments in SAM format, which are then converted to BAM format by using SAMtools [14]. Files with unmap reads are also created.

Both BAM files from read alignment and from input BAM file pre-processing step will be used to generate figure files summarizing mapping information: mapping statistics and mapped read size distribution. In the mapping statistics figure, the proportions of reads with unique and multiple mapping sites in the genome, and unmapped reads are plotted.

5.4 Reads annotation

To find overlaps between read alignments and genomic annotations, intersectBed tool in BEDTools [16] is implemented. Only read alignments which have 100% overlap with annotations are reported by setting -f option in intersectBed to 1.
For reads which can be annotated as given genomic features, detail analyses as shown below will be performed. We also explain the way how we deal with reads with multiple mapping sites and how coordinates of genomic features can be modified.

5.4.1 Reads with multiple mapping sites

A major challenging problem using NGS sequencing data is the annotation of reads aligned at multiple locations. Most of the available frameworks resolve this situation by discarding these reads or by providing random annotations. Here, we propose to keep all the reads aligned to the genome, and to weight them by the number of mapping sites. Suppose a read can be aligned 5 times to the genome, for each mapping site, the read would be counted as 0.2, i.e. 1/5.

5.4.2 Extension parameters

There are four types of extended items which can be used to modify coordinates according to the pattern `\_[iest]\_[+-]Number\_[+-]Number`, as illustrated in Figure 11:

1. **_i_[+-]N1_[+-]N2**: shorten [+-]N1 bp at 5' end, [+-]N2 bp at 3' end
2. **_e_[+-]N1_[+-]N2**: extend [+-]N1 bp at 5' end, [+-]N2 bp at 3' end
3. **_s_[+-]N1_[+-]N2**: get coordinates for sub-region from position N1 to N2 indexed from 5' end
4. **_t_[+-]N1_[+-]N2**: get coordinates for sub-region from position N1 to N2 indexed from 3' end

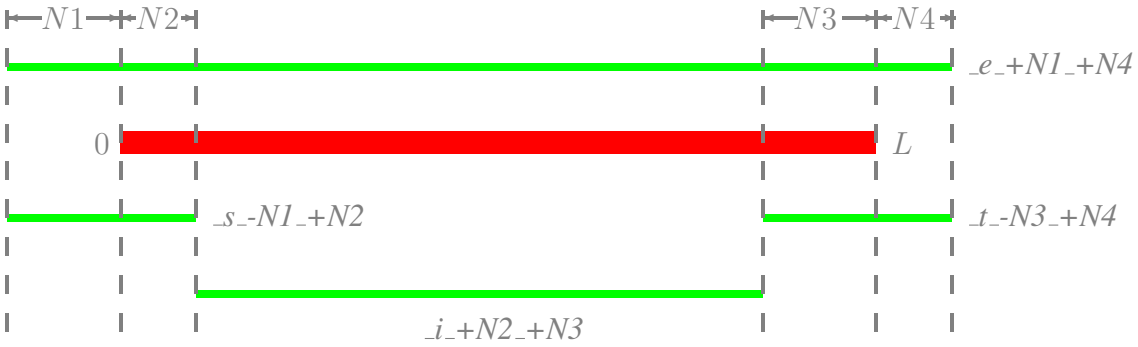

Figure 11: The four types of modification based on the original coordinates

Note that for repeat annotation, extension operations will only perform on the full length repeat in the genome. That means repeat items in the `NCRNA_RMSK_EX` option (see 3.1) if it is specified will

only select untruncated repeats to do the modification and analyses. Repeats representing $>= 90\%$ of its consensus sequence are considered as full length/untruncated repeats.

Here, we show two examples of extension parameters use:

Annotate reads in mature miRNAs. Due to the inaccurate processing of precursor miRNAs by Dicer or downstream miRNA remodelling, mature miRNAs often have end heterogeneities comparing to their annotations in miRBase. Thus, when analyzing mature miRNAs, it is necessary to extend miRNA annotation several bases (e.g. 2 bases) in both upstream and downstream region, which can be easily done in ncPRO-seq 1.6.5 by using **miRNA_e_+2_+2**.

Analyse tRNA-derived small RNAs (tsRNAs). It has been reported that tRNA can be processed again into different types of small RNAs probably through different mechanisms. To check read profiles of these small RNA families, you can specify the following options to **TRNA_UCSC** separated by comma: **tRNA_e_+0_+50** (overview of all tsRNAs), **tRNA_s_+0_+26** (tsRNAs at very 5' end), **tRNA_s_+0_+40** (including tsRNAs from tRNA anti-codon stem cleavage) and **tRNA_t_+0_+50** (tsRNAs from 3' tail of precursor tRNAs).

5.4.3 miRNAs read proportion

In this step, abundant reads mapped in mature miRNA regions are counted, and plotted as the proportion of all mapped reads in the genome. The annotation file of mature miRNA is generated using files from miRBase [12] detailed in 6.3. Each miRNA count is calculated using the intersection of the reads alignement with the mature positions (see intersectBed program from BEDTools).

5.4.4 Reads annotation overview

The reads annotation family is the most general overview, and counts the reads based on the following annotations : coding genes, ncRNAs from Rfam, smallRNAs from repeated regions, rRNAs, and precursor miRNAs from miRBase. One read can belong to several annotations (see multiIntersectBed from BEDTools).

5.4.5 Overview of repeat/Rfam

To compare the read expression in different repeat/Rfam families, we count the number of abundant reads in each family and plot the relative proportion.

Rfam We catalogue non-coding RNA genes in Rfam annotation into five big classes: tRNA, rRNA, snRNA, ACA_snoRNA, CD_snoRNA and others. Note that miRNA annotations are excluded in the Rfam noncoding RNA analyses, for the reason that we have already obtained the miRNA read mapping information in 5.4.3. All rfam annotation files are downloaded from Rfam database [7], please see section 6.1 for more details.

Repeat ncPRO-seq 1.6.5 uses repeat annotations from RepeatMasker [18] results, see section 6.2 for the preparation of the annotation file. We classify different repeats based on the name of repeat family.

5.4.6 Read profiling in each family

Read profiling refers to the analysis of read profiles, which are represented by the distribution of positional read coverage and the read length distribution in annotation family. If the `GROUP_READ` option is activated, for each annotation family, we will compute and plot two types of read profiles, by using abundant and distinct reads respectively, or else only read profiles based on abundant reads will be investigated.

In read profiles, three types of read coverage distribution, which are generated based on 5' end, 3' end and all positions of reads respectively, are displayed together to have a clear view of the biogenesis of small RNAs. Since annotated features belonging to the same single family might have different full length, we use a scaling strategy as shown in Figure 12 to transform read positions in annotated items to corresponding positions in the scaled region. Using this strategy, we are able to sum up positional read coverages from different annotated features in a family, which are then normalized by the number of occurences of each feature position in the genome to obtain the average positional read coverage distribution. For repeat families, additional process is applied before the scaling step that is to locate the annotated repeat regions to the consensus sequence of its family, since repeats from some family like L1 in mouse are always truncated in the genome. Note that the occurrence of read is normalized by the total number of either mapped abundant reads or mapped distinct reads to RPM (reads per million mapped reads) depending on the type of read profiles. We then plot the average read coverage distribution in the scaled region, and number the X-axis according to the median size of annotated items in the family.

In each type of read profiles, relative length distributions of reads mapped in the sense, antisense and both direction of annotation family are plotted.

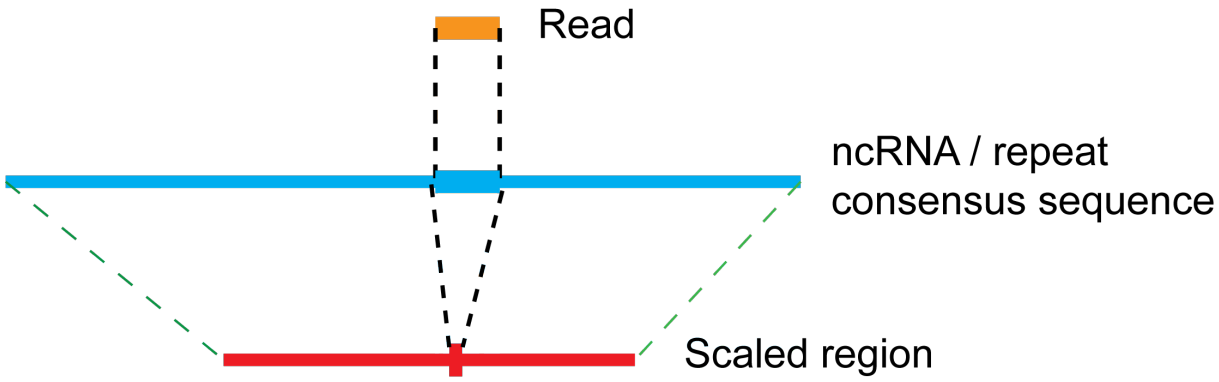

Figure 12: The scaling process

5.4.7 Logo sequences

It is interesting to investigate the base bias in distinct reads from each annotation family, which might give hints about how these small RNAs are processed. In ncPRO-seq 1.6.5, we calculate the fre-

quencies of bases at each position of distinct reads and plot them in sequence logos [3]. Sequence logos can be drawn with respect to 5' end or 3' end of reads depending on the choice of **5** or **3** in `LOGO_DIRECTION` option. Users can choose their favourite way to display sequence logos, either with uniform column heights, or with column heights proportional to informaion content. (`IC_SCALE` option).

For each annotation family, ncPRO-seq 1.6.5 provides two types of sequence logos figures by using different subsets of distinct reads. In one figure, all distinct reads in annotation family are used to create sequence logos, which will give you the processing information of small RNAs like piRNAs that can be produced from various positions in a single region. In another figure, only the distinct read with the highest abundance in each family member is used, which is the case for small RNAs like miRNAs that are accurately processed by enzymes from some special loci.

5.4.8 Read coverage in each annotation item

For each annotation family including different repeat/Rfam families and other custom annotations, ncPRO-seq 1.6.5 creates a table file showing read coverages of each single family member in all sequencing libraries. These table files are compatible with R package like DESeq [1] to identify significantly expressed family members. A global file with all known align reads is also created. For instance, you can find read coverage of each miRNA gene in the miRNA table file. Note that in the tRNA table file, read coverages are computed based on types of tRNA instead of single tRNA genes, since several studies found that expressions of tRNA-derived small RNAs differ from one type of tRNA to another.

5.4.9 Track files

For each sequencing library, for each annotation family, ncPRO-seq 1.6.5 generates compressed track files in both bedgraph and bed formats to describe read mapping in sense and antisense direction of annotation items respectively.

As we have listed in 3.1, there is an option called `GENOME_TRACK_OPTIONS` which allow users to select different subsets of reads mapped in the genome to generate BedGraph formatted track files.

In BedGraph formatted track files, read coverage is computed for each genomic position.

All track files outputed from ncPRO-seq 1.6.5 can be directly uploaded to UCSC browser / other genome browser for visualization.

5.5 Enrichment analysis

ncPRO-seq 1.6.5 has a special engine, which we call **enrichment analysis**, to analyse reads that can not be annotated as known genomic features. Users can select different subsets of reads to perform this analysis by giving different sets of options to `SIG_READ_OPTIONS` (section 6). And reads that can

be aligned to annotation regions given in `EXCLUDE_ANN_GFF` are excluded from enrichment analysis. Finally, we use the following steps to identify regions significantly enriched with remaining reads after filter steps.

1. Slide window of fixed size (`SIG_WIN_SIZE`, e.g. 100,000) along the whole genome at fixed step (`SIG_STEP_SIZE`, e.g. 50,000)
2. For each window, summarize read mapping information (number of mapped reads...)
3. Fit number of mapped reads in all window to selected model to estimate expected read number distribution
4. Compare real and expected read number distribution to determine P value for each window, thereby identify regions with significant numbers of mapped reads (`PVAL_CUTOFF`, e.g. 0.001)
5. Finally generate tables containing read information and additional gene annotation of these regions. Track files in bed format are also created

There are three models that users can choose to do simulation and fit sliding window results:

1. NB.ML: negative binomial distribution inferred using maximum likelihood method
2. NB.012: negative binomial distribution inferred using windows with only 0, 1, or 2 aligned reads
3. Poisson: Poisson distribution inferred using windows with only 0, 1, or 2 aligned reads

For more details about these three models, please check the `addNBSignificance` function in girafe R package [20].

In this step, three types of results are generated: figures displaying the distribution of sliding windows with different number of reads mapped and model simulation results, table files containing location, read mapping and annotation information of identified regions significantly enriched with reads, and track files in bed format.

Note that for small genomes and plant genomes, please choose small sliding window, as big window might lead to the problem of non-existance of window with less than 2 or 10 reads which is important for the model simulation.

6 Annotation files

In ncPRO-seq 1.6.5, we have already genearated noncoding RNA/repeat annotation files for some ($>= 15$) model organisms, including Human,Mouse,Rat, Arabidopsis, Caenorhabditis elegans, Drosophila melanogaster, etc. Please refer to our project website for a complete list of available species.
If users want to create annotation files by themselves in case of working on other speicies or for special usages, please follow the instructions below and refer to readme in annotation/prepareAnnotation folder for commands used to generated these files. Note that ncPRO-seq 1.6.5 only accepts annotation files in gff3 format.

6.1 Rfam annotation

Rfam annotation file can be directly downloaded from Rfam database [7]. Since the reference genomes of some species might changes in different version of Rfam database, make sure that you are using the right Rfam annotation file for your genome assembly. If there is a conflict between the genome assembly used by Rfam and the one you are analyzing, we offer a simple and quick way to solve it by generating new Rfam annotation file for your genome assembly. Basically, we extract Rfam sequences based on the genome annotation in Rfam, then blast them to the new genome, finally use rfam_scan.pl from Rfam database and custom script to create new Rfam annotation file from blast results. And another way to avoid the conflict is: if you have the chain file specifying differences between two different genome assemblies, you can use liftOver from UCSC genome website to get new rfam gff file.

In Rfam annotation file, the value in "Alias" tag in the "attributes" column indicates the ncRNA family information. Example of "attributes" column:

```
ID=RF00026.1;Name=RF00026;Alias=U6;Note=AC157543.8/131368-131259
```

6.2 RepeatMasker annotation

The repeat annotation file is generated from the ouput of RepeatMasker [18]. Users can use our custom scripts to create repeat gff3 file either from UCSC genome browser or from local RepeatMasker output, which are detailed in readme file in annotation/prepareAnnotation folder

There are a list of special feature attributions in repeat annotation files. The "repName","repClass" and "repFamily" features indicate the name, class and family of the repeat respectively. The other three features are used to indicate the location of repeat region in consensus sequence.

Example of "attributes" column:

```
repName=L1_Mur2;repClass=LINE;repFamily=L1;repSt=1413;repEnd=1567;
repFullLen=5877;
```

6.3 miRNA annotation

The miRNA annotation files used in ncPRO-seq 1.6.5 are created based on files from miRBase [12]. The precursor miRNA gff file downloaded from miRBase can be used in ncPRO-seq 1.6.5 after being transformed from gff2 to gff3 format. In miRBase, we cannot find genome coordinates of mature miRNAs. Therefore, we wrote a custom script to generate mature miRNA annotation file based on the mature miRNA sequences, precursor miRNA sequences and precursor miRNA annotation file.

Example of "attributes" column in precursor miRNA annotation file:

`ACC=MI0006363;ID=hsa-mir-1302-2`

Example of ”attributes” column in mature miRNA annotation file:

`Name=mmu-miR-206*;Precursor=mmu-mir-206;ID=mmu.miR.206star`

6.4 tRNA annotation

The tRNA annotation file is generated from tRNA gtf files from UCSC [6] using custom scripts, detailed in readme file in annotation/prepareAnnotation folder. There is a feature in ”attributes” column called ”Type_Name” which indicate the type of tRNA. Example of ”attributes” column:

`ID=chr1.tRNA1547;Type_Name=tRNA-GluTTC`

6.5 piRNA annotation

The piRNA annotation file is generated from piRNA bed files from piRBase [23]. Actually, hg19 and rn5 have piRNA annotation. mm9 piRNA annotation file contains piRNA and piRNA clusters because piRBase gives 47,286,428 piRNA for mm9. It’s too much for a normal use of ncPRO-seq. Clusters are created with this command line : `bedtools merge -s -i piRBase.bed -d 50 -nms`

Example of ”attributes” column in piRNA annotation file (hg19 and rn5):

`Name=piR-rno-26536`

Example of ”attributes” column in piRNA clusters annotation file (mm9):

```
Name=cluster3
Name=piR-mmu-17018147
```

6.6 Other annotations

While ”nc” in the name implies a focus on noncoding RNAs, ncPRO-seq 1.6.5 is far more and can be used to analyse any annotation files in gff3 format like splice site and promoter region of protein coding gene. Note that these annotation files should contain one of the following three features in ”attributes” column to indicate the names of items: Name, Alias or ID.

It has already been reported that small RNAs are enriched at 3’ ends of internal exons (spliRNAs) and at transcription initiation sites (tiRNAs) [19]. To show how the ”other annotations” option works, we create gff3 annotation files of both splice donor site and acceptor site for genomes that has refgene annotation in UCSC [6]. Basically, to generate donor site annotation, we locate the 3’ end of all exons except the last one in genes, and extend 100bp upstream and downstream, thereby obtain regions of size 201bp with 3’ end of exon at position 101. For acceptor site, 5’ end of exons excluding the first

exon in genes are chosen to extend +- 100bp.

Example of ”attributes” column in splice acceptor annotation file:

```
GeneName=NM_001083312;Exon_idx=2;Type=acceptor;Extend_base=100;
```

Example of ”attributes” column in splice donor annotation file:

```
GeneName=NM_001083312;Exon_idx=1;Type=donor;Extend_base=100;
```

7 Installation of additional softwares

We give some simple outlines to install additional softwares required by ncPRO-seq 1.6.5, for more information, please refer to the main site of each software:

- The Bowtie Aligner (<v2.0) [13].
  - The latest Bowtie program can be found here. Users can choose the pre-compiled version for your operating system, or build your own Bowtie from the source (bowtie-*-src.zip). Users need to unzip the downloaded file. If starting with the source, users should use `make` command at the shell prompt to compile the program. In both ways, users should either add the directory where Bowtie binaries are located to your `PATH` variable, or copy/move Bowtie binaries to a directory (e.g. /usr/local/bin) in your `PATH` variable.

    The indexes of the chosen genome reference need to be prepared beforehand and set in the environment variable `BOWTIE_INDEXES`
- The R (>=v3.2.3) [17] and BioConductor [8] softwares.
  - There are several ways to install R program detailed in R-project site. For Mac OS X, the easiest way is to download the latest installer package (R-*.pkg) for Mac OS X, which can be installed by double clicking. For Linux, we just show how to build R from the source. Users can choose a site mirror close to you to download the latest R source code (R-*.tar.gz) for all platforms. You need to unpack the file and go to the unpacked directory at the shell prompt, use

    ```
    > ./configure
    > make
    > su
    > make install
    ```
  - To install Bioconductor, make sure that your computer is connecting to internet and use the following commands in an R command window (type `R` in your shell prompt, and press enter↩):

```
> source("http://bioconductor.org/biocLite.R")
> biocLite()
```

- The R packages: seqLogo [3], girafe [20], RColorBrewer [15], ggplot2 [22], reshape2 [21] and gridExtra [2].

  - To install the seqLogo and girafe package in R, use the following commands in an R command window:

    ```
    > source("http://bioconductor.org/biocLite.R")
    > biocLite("girafe")
    > biocLite("seqLogo")
    ```

  - To install the RColorBrewer, ggplot2, reshape2 and gridExtra packages in R, use the following command in an R command window:

    ```
    > install.packages("RColorBrewer")
    > install.packages("ggplot2")
    > install.packages("reshape2")
    > install.packages("gridExtra")
    ```

- The BEDTools suite (>=v2.15.0) [16] and the bamMapCount program.

  - The latest BEDTools can be downloaded here (BEDTools.*.tar.gz). Use the following commands to compile BEDTools, and copy/move all binaries in bin subdirectory to a directory (e.g. /usr/local/bin) in your PATH variable

    ```
    > tar -zxvf BEDTools.<version>.tar.gz
    > cd BEDTools<version>
    > make clean
    > make all
    > su
    > cp bin/* /usr/local/bin
    ```

  - The bamMapCount program is friendly shared by Assaf Gordon, which can be found here (BEDTools_MapCount_ColorTag.tar.gz). He integrated his programs to BEDTools. Users can unpack the whole package and compile it, then just copy/move the bamMapCount program in bin subdirectory to a directory (e.g. /usr/local/bin) in your PATH variable if you have already installed BEDTools.

    ```
    > tar -zxvf BEDTools_MapCount_ColorTag.tar.gz
    > cd BEDTools_MapCount_ColorTag
    > make all
    > su
    > cp bin/bamMapCount /usr/local/bin
    ```

- The SAMTools suite [14].
  - Users can download the latest version of SAMtools here (samtools-*.tar.bz2). Make sure that you have zlib library (>v1.2.3) installed in your system before compiling SAMtools. Then use the following commands to compile the SAMtools packages, and copy/move 'samtools', 'bcftools/bcftools' and other executables/scripts in 'misc' to a directory (e.g. /usr/local/bin) in your PATH variable:

    ```
    > tar -xvjpf samtools-<version>.tar.bz2
    > cd samtools-<version>
    > make
    > su
    > cp samtools /usr/local/bin
    > cp bcftools/bcftools /usr/local/bin
    ```

- The Convert/ImageMagick utilies.
  - The Convert/ImageMagick normally comes with most of modern linux/UNIX systems, but if you also want to install another version, there are a few easy steps to do which are detailed here. For Mac OS X, users can use MacPorts command as shown below to build ImageMagick in your system.

    ```
    > su
    > port install ImageMagick
    ```

- The Python package: ReportLab.
  - To install ReportLab package in Python (<v2.7): ReportLab can be downloaded here. Use the following commands to compile ReportLab:

    ```
    > tar -zxvf reportlab-2.7.tar.gz
    > cd reportlab-2.7
    > python setup.py install --user
    ```

  - To install ReportLab package in Python (>=v2.7): Use the following commands to compile ReportLab:

    ```
    > easy_install reportlab

    OR

    > pip install reportlab
    ```

- A local server with php.
  - For macOS users, most of the time, Apache is already installed but not activated. The different steps to enabled Apache and php are described here. For other linux/UNIX users, if the webserver is not installed, XAMPP might be a good choice.

8 ncPRO-seq on Galaxy

You can install ncPRO-seq on your Galaxy [11] [4] [10] instance.

8.1 ncPRO-seq dependencie and Galaxy settings

To install ncPRO-seq on your Galaxy instance, you must install Docker tool. It's the only dependencie !

The Docker program:

- For Linux users : install Docker for Linux
- For MacOS users : install Docker for MacOS
- For Windows users : install Docker for Windows

Galaxy settings to use Docker images:

- To display Tool Shed tools, you must change a parameter in `galaxy/config/galaxy.ini`

  ```
  tool_dependency_dir = ../tool_dependencies
  ```

- To use Docker images, add new "destination" tag in `galaxy/config/job_conf.xml`

```
<?xml version="1.0"?>
<!-- A sample job config that explicitly configures job running
the way it is configured by default (if there is no explicit config). -->
<job_conf>
    <plugins>
        <plugin id="local" type="runner"
load="galaxy.jobs.runners.local:LocalJobRunner" workers="4"/>
    </plugins>
    <handlers>
        <handler id="main"/>
    </handlers>
    <destinations default="docker_local">
       <destination id="local" runner="local"/>
       <destination id="docker_local" runner="local">
            <param id="docker_enabled">true</param>
            <param id="docker_memory">2G</param>
            <param id="docker_sudo">false</param>
            <param id="docker_net">bridge</param>
       </destination>
    </destinations>
</job_conf>
```

8.2 ncPRO-seq installation with Galaxy Tool Shed

- Go in **Admin** section
- In **Tools and Tool Shed** part, click on **Search Tool Shed**
- Click on **Galaxy Main Tool Shed**
- In search bar, write **ncproseq** and press enter↩
- Select ncPRO-seq 1.6.5 and **Preview and install**
- Click on **Install to Galaxy**
- Add new tool panel section or select existing tool panel section and click on **Install**

When "Status" are **Installed** (refresh page), ncPRO-seq tools are on your Galaxy instance. You have three tools: Alignment and QC, Annotation and Profiling.

8.3 ncPRO-seq tools

8.3.1 Alignment and QC

Input : fastq or bam file(s) (maximum: 4)
Output :

- report (html, pdf or both)
- log file
- optional : bam file (annotation overview using the RFAM and RepeatMasker database)

ncPRO-seq steps realized by Alignment and QC tool:

Alignment and QC steps	Description
processRead (if input file is a fastq file)	calculate read length distribution, median quality score for each postion, and group reads
processBam (if input file is a bam file)	processed and group reads from bam files
mapGenome (if Run Alignment = Yes)	run Bowtie for genome mapping
mapGenomeStat (if Run Alignment = Yes)	compute number of mapped reads and unmapped reads in the genome
mapAnnOverview (if Run Alignment = Yes)	compute overview of reads annotation
overviewRmsk (if Generate the annotation overview = Yes)	compute read coverage for each repeats family
overviewRfam (if Generate the annotation overview = Yes)	compute read coverage for each ncRNA family
html_builder	build the html report file
pdf_builder	build the pdf report file

Table 4: Description of Alignment and QC steps

8.3.2 Annotation

Input : bam file
Output :

- annotation file (RPM normalization or not)
- log file
- optional : annotation file with all miRBase miRNA (RPM normalization or not)
- optional : IGV tracks for ncRNA visualisation

ncPRO-seq steps realized by Annotation tool:

Annotation steps	Description
processBam (if bam file isn't grouped)	processed and group reads from bam files
generateNcgff	create gff file for special ncrna family
ncrnaProcess	ncRNA family analyses, including read coverage, read length distribution, read coverage in subfamilies, and sequence logo
ncrnaTracks (if Create IGV tracks = Yes)	genreate ucsc tracks for ncrna visualization

Table 5: Description of Annotation steps

8.3.3 Profiling

Input : bam file
Output :

- png of the annotation profiling (distinct reads)
- png of the annotation profiling (abundant reads)
- log file

ncPRO-seq steps realized by Profiling tool:

Profiling steps	Description
processBam (if bam file isn't grouped)	processed and group reads from bam files
generateNcgff	create gff file for special ncrna family

ncrnaProcess	ncRNA family analyses, including read coverage, read length distribution, read coverage in subfamilies, and sequence logo

Table 6: Description of Profiling steps

9 FAQ

9.1 Enabling CGI using XAMPP

If not well configured, the server can send a "Error 403 Access forbiden" during the CGI execution. This could happen for instance with the file browser system of ncPRO-seq 1.6.5. To correct the error, be sure that the httpd.conf file of XAMPP is well configured. For instance, on MacOS, edit the file /Application/XAMPP/etc/httpd.conf and add the following parameters.

```
<Directory "/Applications/XAMPP/xamppfiles/cgi-bin/">
    AllowOverride None
    Options Indexes FollowSymLinks MultiViews ExecCGI
    Order allow,deny
    Allow from all
</Directory>
```

This will enabling the CGI to be executed.

9.2 Working with a new genome

In ncPRO-seq 1.6.5, it is possible to do small RNA profiling analyses in a new genome without giving any known annotations (miRNA, tRNA, etc). For example, you want to work in yeast. What you need to do is:

1. create the bowtie index of yeast genome with index name "yeast", put them in your Bowtie index folder

2. create a folder "yeast" in MY_INSTALLATION_DIR/annotation directory

3. create the "chrom.sizes" file of your yeast genome in MY_INSTALLATION_DIR/annotation/yeast folder. "chrom.sizes" is the reference idex file, which has two columns (name of chromosome and size of the chromosome) and can be generated using samtools

9.3 other FAQs

Please refer to http://ncpro.curie.fr/faq.

References

[1] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11:R106, 2010.

[2] Baptiste Auguie. *gridExtra: functions in Grid graphics*, 2012. R package version 0.9.1.

[3] Oliver Bembom. *seqLogo: Sequence logos for DNA sequence alignments*. R package version 1.18.0.

[4] Daniel Blankenberg, Gregory Von Kuster, Nathaniel Coraor, Guruprasad Ananda, Ross Lazarus, Mary Mangan, Anton Nekrutenko, and James Taylor. Galaxy: A web-based genome analysis tool for experimentalists. *Current protocols in molecular biology*, pages 19–10, 2010.

[5] P. Brodersen and O. Voinnet. The diversity of RNA silencing pathways in plants. *Trends Genet.*, 22:268–280, May 2006.

[6] T. R. Dreszer, D. Karolchik, A. S. Zweig, A. S. Hinrichs, B. J. Raney, R. M. Kuhn, L. R. Meyer, M. Wong, C. A. Sloan, K. R. Rosenbloom, G. Roe, B. Rhead, A. Pohl, V. S. Malladi, C. H. Li, K. Learned, V. Kirkup, F. Hsu, R. A. Harte, L. Guruvadoo, M. Goldman, B. M. Giardine, P. A. Fujita, M. Diekhans, M. S. Cline, H. Clawson, G. P. Barber, D. Haussler, and W. James Kent. The UCSC Genome Browser database: extensions and updates 2011. *Nucleic Acids Res.*, 40:D918–923, Jan 2012.

[7] P. P. Gardner, J. Daub, J. Tate, B. L. Moore, I. H. Osuch, S. Griffiths-Jones, R. D. Finn, E. P. Nawrocki, D. L. Kolbe, S. R. Eddy, and A. Bateman. Rfam: Wikipedia, clans and the "decimal" release. *Nucleic Acids Res.*, 39:D141–145, Jan 2011.

[8] Robert C Gentleman, Vincent J. Carey, Douglas M. Bates, and others. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80, 2004.

[9] M. Ghildiyal and P. D. Zamore. Small silencing RNAs: an expanding universe. *Nat. Rev. Genet.*, 10:94–108, Feb 2009.

[10] Belinda Giardine, Cathy Riemer, Ross C Hardison, Richard Burhans, Laura Elnitski, Prachi Shah, Yi Zhang, Daniel Blankenberg, Istvan Albert, James Taylor, Webb C Miller, W James Kent, and Anton Nekrutenko. Galaxy: a platform for interactive large-scale genome analysis. *Genome research*, 15(10):1451–1455, 2005.

[11] Jeremy Goecks, Anton Nekrutenko, James Taylor, and The Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*, 11(8):R86, 2010.

[12] A. Kozomara and S. Griffiths-Jones. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*, 39:D152–157, Jan 2011.

[13] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, 10:R25, 2009.

[14] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25:2078–2079, Aug 2009.

[15] Erich Neuwirth. *RColorBrewer: ColorBrewer palettes*, 2011. R package version 1.0-5.

[16] A. R. Quinlan and I. M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26:841–842, Mar 2010.

[17] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

[18] AFA. Smit and R. Hubley. RepeatModeler Open-1.0. *http://www.repeatmasker.org*, 2008-2010.

[19] R. J. Taft, C. Simons, S. Nahkuri, H. Oey, D. J. Korbie, T. R. Mercer, J. Holst, W. Ritchie, J. J. Wong, J. E. Rasko, D. S. Rokhsar, B. M. Degnan, and J. S. Mattick. Nuclear-localized tiny RNAs are associated with transcription initiation and splice sites in metazoans. *Nat. Struct. Mol. Biol.*, 17:1030–1034, Aug 2010.

[20] Joern Toedling, Constance Ciaudo, Olivier Voinnet, Edith Heard, and Emmanuel Barillot. girafe - an R/Bioconductor package for functional exploration of aligned next-generation sequencing reads. *Bioinformatics*, 26:2902–2903, 2010.

[21] Hadley Wickham. Reshaping data with the reshape package. *Journal of Statistical Software*, 21(12):1–20, 2007.

[22] Hadley Wickham. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009.

[23] Peng Zhang, Xiaohui Si, Geir SkogerbÃ¸, Jiajia Wang, Dongya Cui, Yongxing Li, Xubin Sun, Li Liu, Baofa Sun, Runsheng Chen, Shunmin He, and Da-Wei Huang. pirbase: a web resource assisting pirna functional study. *Database*, 2014, 2014.