ncPRO-seq

Annotation and profiling of ncRNAs from sRNA-seq data

Chong-Jian Chen [1-5,†] Nicolas Servant [1,4,5,†] , Joern Toedling [1-6] , Alexis Sarazin[7] , Antonin Marchais [8] , Evelyne Duvernois-Berthet [7] , Valerie Cognat [9] , Vincent Colot [7] , Olivier Voinnet [8] , Edith Heard [1-3,‡] , Constance Ciaudo [1,2,3,8,‡] and Emmanuel Barillot [1,4,5,‡]

[1] Institut Curie, Paris, France [2] CNRS UMR3215, Paris, France [3] INSERM U934, Paris, France [4] INSERM U900, Paris, France, [5] Mines ParisTech, Fontainebleau, France, [6] Institute of Molecular Biology gGmbH, Mainz, Germany [7] Institut de Biologie de l'Ecole Normale Supérieure, CNRS UMR8197, INSERM U1024, Paris, France, [8] Swiss Federal Institute of Technology Zurich, Department of Biology, Chair of RNA biology, Zurich, Switzerland, [9] Institut de Biologie Moléculaire des Plantes, CNRS UPR2357, Université de Strasbourg, Strasbourg, France

[†] contributed equally to this work, [‡] co-last author

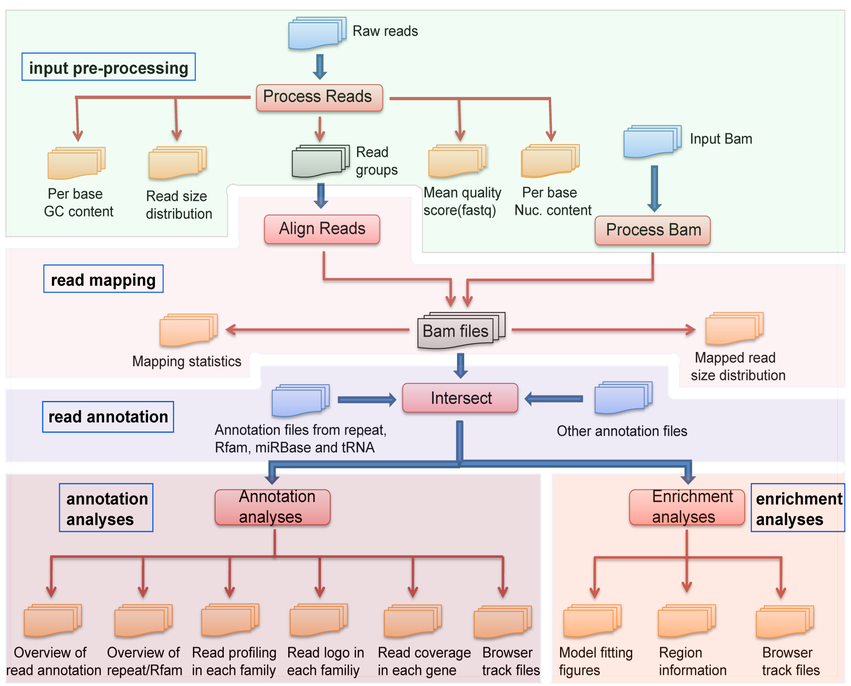
bioinfo.ncproseq@curie.fr

Over recent years, deep sequencing technology has become a powerful approach for 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 [1]. Moreover, in other organisms like plants more classes of small ncRNA have been described indicating that a wide range of small ncRNAs exist [2].

However, most of the existing tools devoted to sRNA-seq analysis, are only based on miRNAs annotation and quantification, significantly neglecting other types or new types of small ncRNAs. Moreover, they just perform gene-based analysis, but not detailed family-based (profiling) analysis which is critically important to investigate known small ncRNA families and to identify novel small ncRNA families.

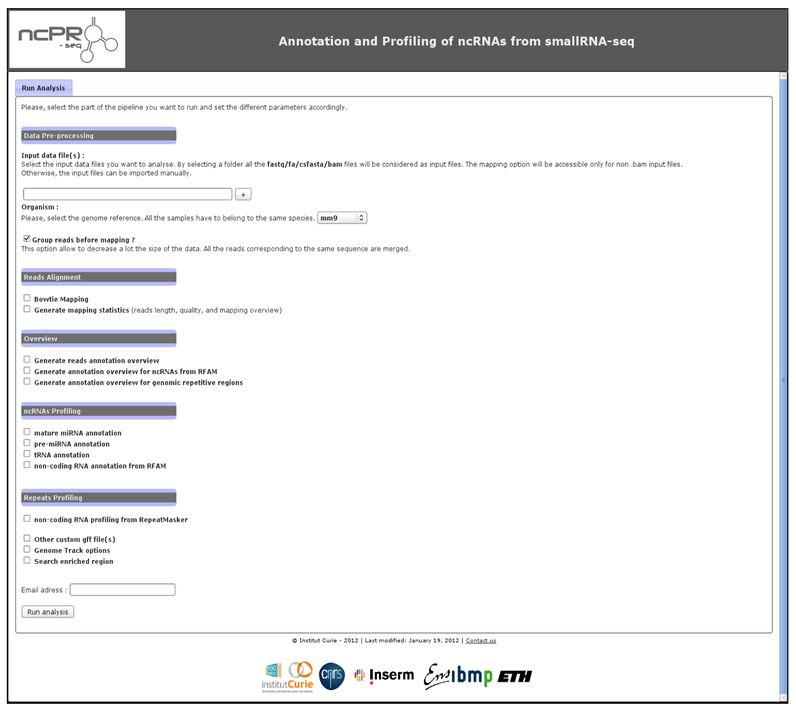
Here we present a **comprehensive** and **flexible** ncRNA analysis pipeline, **ncPRO-seq** (**N**on-**C**oding RNA **PRO**filing from sRNA-seq), which is able to interrogate and perform detailed profiling analysis on small RNAs derived from annotated non-coding regions in **miRBase**, **Rfam** and **repeatMasker**, as well as regions defined by users. We perform both **gene-based** and **family-based** detailed analyses of small RNAs. The ncPRO-seq pipeline also has a module to identify regions significantly enriched with short reads that can not be classified as known ncRNA families [3], thus enabling the discovery of yet unknown 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. Finaly, the ncPRO-seq pipeline can be used to analyze data based on genome from metazoan to plants. The current version proposes annotation files for **fifteen** different species.

ncPRO-seq Workflow

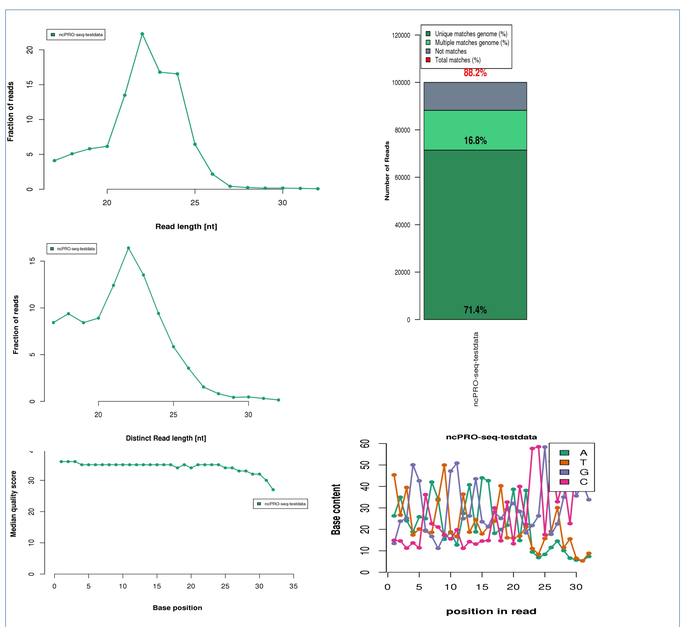


- Support multiple Solexa, SOLiD, 454 raw reads, and Bam files
- Reads grouping strategy (distinct vs abundant reads)
- Quality control of raw and aligned reads
- Reads mapping using the Bowtie software [4]
- More than 15 annotated organisms from mammals/ metazoan to plants
- Flexible annotation and analysis of ncRNA families from Rfam, UCSC tRNA and miRBase
- Annotation and analysis of repeats classes from RepeatMasker
- Support user defined annotation files (gff3)
- Detect regions significantly enriched with reads
- Settings of UCSC Genome Browser tracks for visualization
- Stand-alone/command line pipeline and user-friendly interface

Running ncPRO-seq

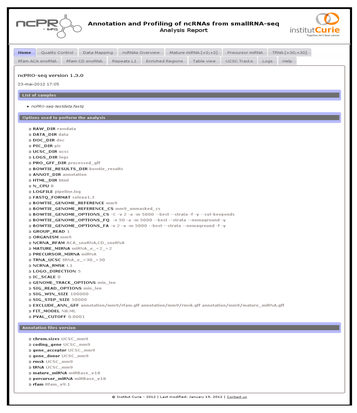


Quality Control and mapping



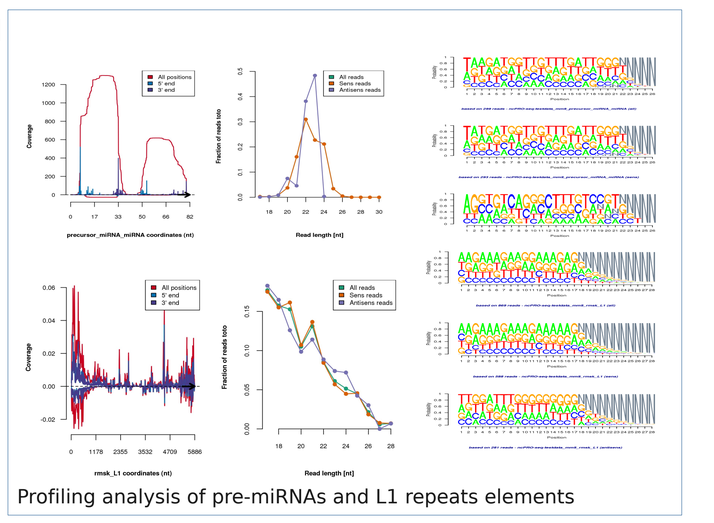
Quality Control (insert size, base composition and quality, mapping, etc.)

ncPRO-seq Interactive Analysis Report



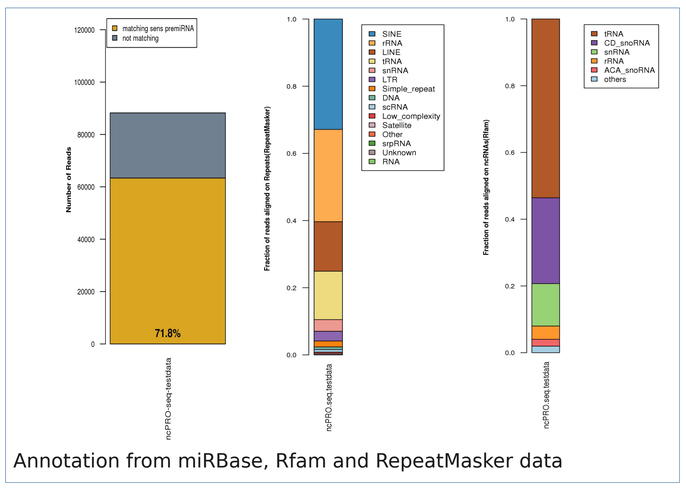
Export and Visualization

Family-based Analysis



Profiling analysis of pre-miRNAs and L1 repeats elements

Annotation of ncRNAs and repeats classes



Annotation from miRBase, Rfam and RepeatMasker data

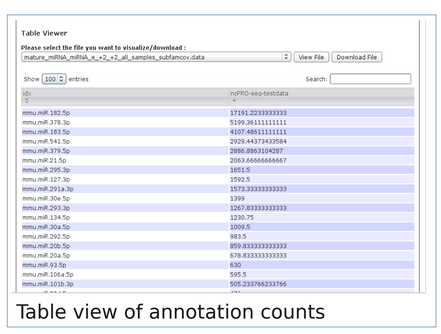
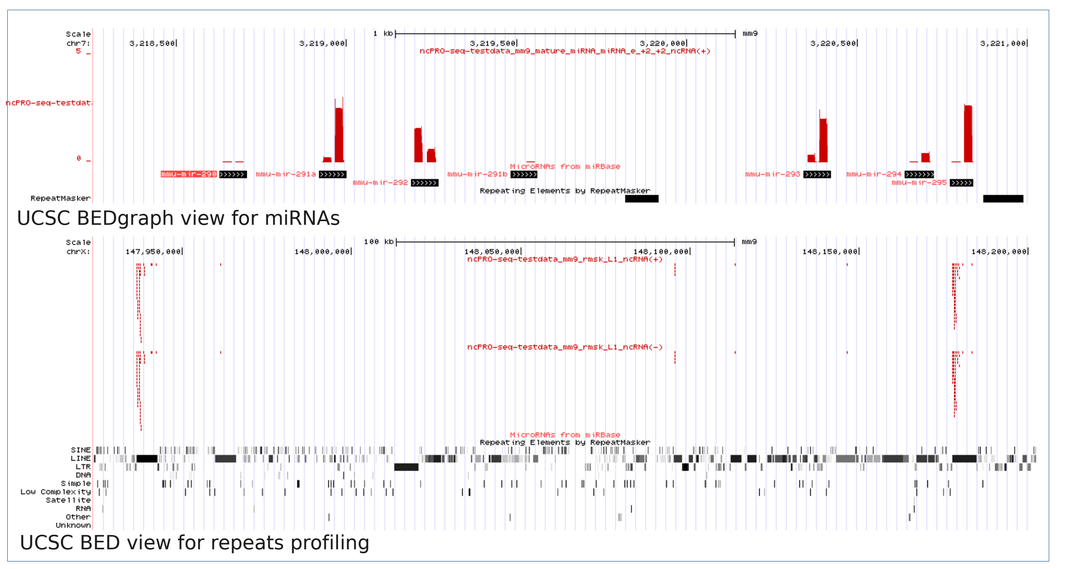


Table view of annotation counts



UCSC BEDgraph view for miRNAs

UCSC BED view for repeats profiling

The ncPRO-seq pipeline is a stand-alone pipeline, which can be easily installed in a local computer or cluster. We offer two ways to launch the pipeline, through either a command line or a user-friendly web interface. The ncPRO-seq pipeline allows users to specify different options at each analysis stage, from raw reads processing to ways to generate results, all of which can be done by either selecting parameters in the web page or manually editing a configuration file. The results are available through an HTML report. Users can directly view figures and tables in the result web page. Track files are generated for visualization in genome browsers.

We deploy the ncPRO-seq pipeline in **http://ncproseq.sourceforge.net**, where users can find detailed information, such as basic descriptions, manuals, test dataset and example results. An online version for small dataset is available at **http://ncproseq.sourceforge.net/online.html**

[1] M. Ghildiyal, PD. Zamore. Small silencing RNAs: an expanding universe. Nat Rev Genet., 10(2):94-108, 2009.
[2] P. Brodersen, O. Voinnet. The diversity of RNA silencing pathways in plants. Trends Genet., 22(5):268-80, 2006
[3] J. Toedling, C. Ciaudo, O. Voinnet, E. Heard, and E. Barillot. Girafe–an R/Bioconductor package for functional exploration of aligned next-generation sequencing reads.Bioinformatics, 26, 2902–2903, 2010.
[4] B. Langmead, C. Trapnell, et al. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol., 10, R25.

institutCurie Ensemble, prenons le cancer de vitesse. · cnrs · Inserm · Ens · ibmp · ETH Eidgenössische Technische Hochschule Zürich Swiss Federal Institute of Technology Zurich