**Title:** Characterization of the small RNA transcriptome using the bcbio-nextgen python framework

**Authors:** Lorena Pantano, Brad Chapman, Rory Kirchner, John Hutchinson, Oliver Hofmann, Shannan Ho Sui
**Affiliations** Harvard T.H. Chan School of Public Health, Boston. Wolfson Wohl Cancer Research Centre
**Contact:** lpantano@hsph.harvard.edu

**Project:** https://github.com/chapmanb/bcbio-nextgen


The study of small RNA helps us understand some of the complexity of gene regulation of a cell. Of the different types of small RNAs, the most important in mammals are miRNA, tRNA fragments and piRNAs. The advantage of small RNA-seq analysis is that we can study all small RNA types simultaneously, with the potential to detect novel small RNAs. bcbio-nextgen is a community-developed Python framework that implements best practices for next-generation sequence data analysis and uses gold standard data for validation. We have extended bcbio to include a small RNA-seq analysis pipeline that performs quality control, removal of adapter contamination, annotation of miRNA, isomiRs and tRNAs, novel miRNA discovery, and genome-wide characterization of other types of small RNAs. The pipeline integrates tools such as miRDeep2[1], seqbuster[2], seqcluster[3] and tdrMapper[4] to facilitate annotation to small RNA categories. It produces a R Markdown template that helps with downstream statistical analyses in R, including quality control metrics and best practices for differential expression and clustering analyses. Finally, the pipeline generates an interactive HTML-based browser for visualization purposes. This is useful for characterizing novel small RNA types, working with non-model organisms, or providing a general profiling description. This browser shows the small RNA regions along with their genomic annotation, expression profile over the precursor, secondary structure, and the top expressed sequences. Here, we show the capabilities of the pipeline and validation using data from the miRQC project. We show that the quantification accuracy is around 95% for miRNAs. We obtained similar results for other types of small RNA molecules, demonstrating that we can reliably detect small RNAs without a dependency on specific databases.

[1] Friedlander, M. R., MacKowiak, S. D., Li, N., Chen, W., & Rajewsky, N. (2012). MiRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Research, 40(1), 37–52.

[2] Pantano, L., Estivill, X., & Martí, E. (2010). SeqBuster, a bioinformatic tool for the processing and analysis of small RNAs datasets, reveals ubiquitous miRNA modifications in human embryonic cells. Nucleic Acids Research, 38(5), e34

[3] Pantano, L., Friedlander, M. R., Escaramis, G., Lizano, E., Pallares-Albanell,

J., Ferrer, I., ... Marti, E. (2015). Specific small-RNA signatures in the amygdala at premotor and motor stages of Parkinson's disease revealed by deep sequencing analysis. Bioinformatics (Oxford, England).

[4] Selitsky, S. R., & Sethupathy, P. (2015). tDRmapper: challenges and solutions to mapping, naming, and quantifying tRNA-derived RNAs from human small RNA-sequencing data. BMC Bioinformatics, 16(1), 354.