smallRNA data analysis

Angelica Tulipano, Andreas Gisel
Institute for Biomedical Technologies – CNR, Bari, Italy
http://www.ba.itb.cnr.it/

The discovery of small RNA, such as miRNA and siRNA, opened a new dimension in gene regulation on the level of transcriptional and post-transcriptional regulation (1). To understand the distribution and expression levels of small RNAs is therefore crucial to understand tissue development (2), diseases (3), therapies with xenobiotic medicaments (4) or with small RNAs (5). Furthermore, each cell type, each tissue has a different onset of small RNAs and their expression. Only a large amount of samples from all these tissues will reveal the whole "small RNA-om". Technologies such as NGS heavily supports the investigations of these small RNA such as that a deep sequencing approach gives a holistic view of a snapshot of the small RNA regulatory activity in a biological sample. With the increasing number of sequence output offered by the different NGS technologies, the analysis of these large numbers of sequences especially for small RNA data analysis become time consuming and prone of errors in respect of the prediction of new small RNAs.

Because NGS produces in one experiment such a large number of sequences the technologies offer to run in parallel several samples labelled with a short barcode sequences. Therefore a typical workflow to analyse such a deep sequencing small RNA data set starts with the identification of these barcodes at the 5' end of the reads from up to 100 million sequences, remove the barcode sequence and search at the 3' end for the adaptor sequence and remove also these sequence fragments; logistically not too complex but computational very intensive. An intelligent distribution on different threads per CPU, on a GPU, in a cloud or over the GRID would dramatically reduce this process. The next step is the mapping of these cleaned reads onto the reference genome and find potential precursor sequences from known or new miRNA genes which would fold in a proper stem-loop secondary structure. This second more complex step is also computational demanding but more important includes a process for the selection of the proper folding for the cutting site to produce the mature miRNA and the miRNA. Since the feature of such a folding is quite broad the workflow needs to be flexible and user controllable to adjust a range of parameter to extract a list of significant potential miRNA genes and the corresponding mature miRNA.

We are developing a workflow, which starts with the read processing from multiplexed sequencing data (Illumina) and offers a mapping procedure and a corresponding miRNA recognizing procedure with a range of parameters to adjust the output.

References

1. Taft, R. J., Pang, K. C., Mercer, T. R., Dinger, M., & Mattick, J. S. (2010). Non-coding RNAs: regulators of disease. The Journal of pathology, 220(2), 126–139. doi:10.1002/path.2638.
2. Morin, R. D., O'Connor, M. D., Griffith, M., Kuchenbauer, F., Delaney, A., Prabhu, A.-L., Zhao, Y., et al. (2008). Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. Genome Research, 18(4), 610–621. doi:10.1101/gr.7179508.
3. Joyce, C. E., Zhou, X., Xia, J., Ryan, C., Thrash, B., Menter, A., Zhang, W., et al. (2011). Deep sequencing of small RNAs from human skin reveals major alterations in the psoriasis miRNAome Human molecular genetics, 20(20), 4025–4040. doi:10.1093/hmg/ddr331.
4. Rodrigues, A. C., Li, X., Radecki, L., Pan, Y.-Z., Winter, J. C., Huang, M., & Yu, A.-M. (2011). MicroRNA expression is differentially altered by xenobiotic drugs in different human cell lines Biopharmaceutics & drug disposition, 32(6), 355–367. doi:10.1002/bdd.764.
5. Gandellini, P., Profumo, V., Folini, M., & Zaffaroni, N. (2011). MicroRNAs as new therapeutic targets and tools in cancer. Expert opinion on therapeutic targets, 15(3), 265–279. doi:10.1517/14728222.2011.550878.