A combinatorial and integrated method to analyse RNA-seq reads

Nicolas Philippe, Mikael Salson, Thérèse Commes, Eric Rivals

A combinatorial and integrated method to analyse RNA-seq reads

Nicolas Philippe, Mikael Salson, Therese Commes, Eric Rivals
Laboratoire d'Informatique, de Robotique et de Microélectronique, UMR 5506 CNRS,
Université de Montpellier II, Montpellier, France;
LIFL, CNRS, INRIA Lille, Univ. Lille I, Villeneuve d'Ascq, France;
CRBM, UMR 5237 CNRS, Montpellier, France
http://www.lirmm.fr/~rivals

RNA sequencing enables a complete investigation covering the full dynamic spectrum of a transcriptome. It thus paves the way to a better understanding of the function of gene expression in different tissues, during development or pathological states. However, the splicing process, which generates both co-linear and non co-linear RNAs, the inclusion of sequencing errors, somatic mutations, polymorphisms, and rearrangements make the reads differ from the reference genome in a variety of ways. This complicates the task of comparing reads with a genome. Currently, the analysis paradigm consists in:

1. mapping the reads to a reference genome contiguously allowing as many differences as one expects to be necessary to accommodate sequence errors and small polymorphisms;

2. using uniquely mapped reads to determine covered genomic regions, either for computing a local coverage to predict mutations and filter out sequence errors (cf. program ERANGE), or for delimiting expressed exons approximately (cf. program TopHat);

3. re-aligning unmapped reads, which were not mapped contiguously at step one, to reveal splicing junctions.

Limitations of this approach include lack of precision, redundant computations due to multi-mapping steps, error propagation due to heuristics and the absence of back-tracking. We propose a novel, integrated approach to analyze nowadays longer reads (> 50 bp). The idea is to adopt a k-mer approach that combines the genomic positions and local coverage to perform a complex analysis of each read and detect in a single step, mutations, indels, errors, as well as both normal and chimeric splice junctions. Comparisons with other tools demonstrate the feasibility of this approach, which yields both sensitive and highly specific inferences.

References

1. N. Philippe, M. Salson, T. Lecroq, M. Leonard, T. Commes and E. Rivals; Querying large read collections in main memory: a versatile data structure. BMC Bioinformatics, Vol. 12, p. 42, doi:10.1186/1471-2105-12-242, 2011.

Relevant Web sites

2. http://crac.gforge.inria.fr/gkarrays/
3. http://www.atgc-montpellier.fr/ngs/