Séminaire – Math For Genomics

Séance du jeudi 18 novembre 2021. 14h.
Evry. IBGBI. Grand Amphi

RNA-Seq in plant organelles: why, how and new statistical approach

Benoît Castandet (IPS2 - Université Paris-Diderot)

Development of RNA-Seq strategies to study the chloroplast transcriptome

Organelles are descendants of ancient, free-living bacteria, and during the course of evolution, most of the endosymbiont genetic content has been lost. Expression of the remaining genes now depends on a complex interplay of nuclear and organelle-encoded proteins. As many as three RNA polymerases and six sigma factors create the primary transcriptome, which is further processed by an array of ribonucleases, RNA-binding proteins, and RNA splicing and editing factors, ultimately producing mature transfer RNAs, ribosomal RNAs, and messenger RNAs (tRNAs,rRNAs, and mRNAs, respectively). As a consequence, deciphering the roles of these various factors in post-transcriptional maturation in vivo is tedious and labor intensive, with the extensive use of traditional molecular biology tools like RNA blots. This has long impaired any global understanding of the organellar RNA maturation processes. These limitations have been largely overcome by the wide adoption of high-throughput cDNA sequencing (RNA-Seq) for the last ten years. In particular, combining the power of RNA-Seq to the precision of in-depth molecular analysis proved to be an invaluable methodology to study RNA editing, splicing and to decipher the role of chloroplast ribonucleases.

Arnaud Liehrmann (IPS2 - Université d'Évry)

Automatic differential analysis of transcription variants in the chloroplast with change-point detection analysis

Several RNA-Seq based strategies have recently been developed to decipher the chloroplast gene expression complexity. Most of the tools developed, however, only count the abundance of sequencing reads along annotated patterns (typically genes) and therefore neglect non-coding regions and regulatory events within genes that are pervasive in the chloroplast transcriptome. In the context of differential expression analysis, these events result in local changes in the log-ratio of coverage along the genome between compared conditions. Our method, DiffSegR, allows systematic identification of differential maturation events without relying on pre-existing annotations in a two-step design: (1) *Summary of the transcriptionnal landscape.* We propose to identify these local changes, also called changepoints, in the log-ratio of RNA-Seq signals using a fast multiple changepoints detection algorithm that optimizes a penalized likelihood criteria [4]. These changepoints define the limits of intervals within which overlapping reads are then summarized. The first step ends by building a count matrix. (2) *Differential expression analysis.* Each Interval, through its associated row in the count matrix, is statistically assessed for quantitative changes in expression levels between compared conditions using the negative binomial model of edgeR [5] or DESeq2 [3]. We illustrate our method on two RNA-Seq datasets that were previously used in combination with traditional molecular biology techniques to decipher the role of the chloroplast ribonucleases PNPase [1] and MiniIII [2]. On these two sets of maturation events, DiffSegR returns results close to the expert annotation of the chloroplast RNA-Seq signals performed by biologists, while reducing the analysis time from several hours to a few minutes.

References

[1] Benoît Castandet, Amber M. Hotto, Zhangjun Fei, and David B. Stern. Strand-specific rna sequencing uncovers chloroplast ribonuclease functions. *FEBS Letters*, 587(18):3096–3101, 2013.

[2] Amber M. Hotto, Benoît Castandet, Laetitia Gilet, Andrea Higdon, Ciarán Condon, and David B. Stern. Arabidopsis Chloroplast Mini-Ribonuclease III Participates in rRNA Maturation and Intron Recycling. *The Plant Cell*, 27(3):724–740, 2015.

[3] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. 15(12), December 2014.

[4] Robert Maidstone, Toby Hocking, Guillem Rigaill, and Paul Fearnhead. On optimal multiple changepoint algorithms for large data. *Statistics and Computing*, 27:1573–1375, 2017.

[5] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. 26(1):139–140, November 2009.