M2 Mycologie

Analyses RNA-Seq

Université
Paris Cité

Principe des analyses RNA-Seq

- revues de bonnes pratiques : [2, 5]
- outils statistiques : Bioconductor ; [1, 3]

Etapes d'analyse

1. Extraction ADN à partir d'un échantillon
2. DNA sequencing
3. [*] Alignement des "reads" sur un transcriptome
4. Analyse exploratoire des données (contrôle qualité, couverture, etc.)
5. Identification des variants (SNP, indels – small insertions and deletions)
6. Quantification de gènes (statistiques sur des données de comptage)

Principes de l'analyse

Les "analyses NGS" (RNA, CHIP, etc.) doivent prendre en compte une estimation de la variance intra-groupe lors de l'analyse de multiples gènes, d'où l'idée de combiner l'information entre les gènes. L'approche DESeq permet de détecter et corriger les estimations de dispersion qui restent trop faibles en modélisant la dépendance entre la dispersion de l'expression moyenne sur l'ensemble des échantillons. De plus, cette approche fournit une méthode originale de classement des gènes et de visualisation des tailles d'effet [4]. La librairie R DESeq2 fournit également des estimateurs de type "shrunken fold changes" (avec leur erreur-type).

TopHat + HTSeq (or FeatureCounts)

Cette approche, disponible sous Galaxy, prend plus de temps et repose sur le logiciel TopHat. Celui-ci présente l'avantage de rendre compte des jonctions intron-exon, ce qui le distingue de bowtie.

Kallisto (ultra-fast mapping)

- Entrée/Sortie = SRA ou Fastq et fichier d'annotation[1] -> TPM
- plus rapide que TopHat, mais moins précis
- chaque "run" doit être en triplicat

[1]Attention : l'ID du transcrit doit correspondre exactement à l'ID du gène dans le fichier GFF3 d'annotation.

Analyse des données de comptage

- contrôle qualité (nombre de hits trop bas)
- classification automatique et analyse en composantes principales
- analyse différentielle

Analyse des données de comptage (2)

Normalization refers to the comparison of size factors defined as the median of ratios of each sample to a virtual reference sample (median of each gene's values across samples). Those ratios are expected to match the ratios of the library sizes and be roughly equal to one. Dividing each column of the count table by the corresponding size factor yields normalized count values, which can be scaled to give a counts per million interpretation. Note that this is different from the approach taken by `edgeR`, which considers the trimmed mean of M values.

The MA plot shows the average vs mean–difference of log fold change, centered around 0, and it is expected to observe higher variability of log ratios at lower counts.

Analyse des données de comptage (3)

A PCA or simple heatmap of the results of hierarchical clustering of the sample data can be used to assess overall similarity between samples. We expect triplicate to cluster together while samples from very different experimental conditions are expected to be far away one from the other. Of note, it is useful to apply a regularized log–transformation on the raw counts to avoid the impact of few highly variale genes, hence considering a roughly equal contribution from all genes. For genes with high counts, this mostly resembles a log2 transformation, whereas for genes with low counts, this shrunkens values toward gene's average across amples. It is important to note that this is only for exploratory analysis; for statistical modeling, raw counts should be preferred since DESeq will handle appropriate correction automagically.

Analyse différentielle

The statistical model underlying differential analysis of count data is a Negative Binomial, which contrary to the standard Poisson model allows to account for overdispersion (i.e., variance greater than mean). Variance is modeled as $\mathbb{V}[NB(\mu, \alpha)] = \mu + \alpha\mu^2$, and the very first step in differential analysis is to get an estimate of the dispersion parameter for each gene (independent of the condition, which is sensible since there is usually a low number of replicates).

Notice that for genes with very low read counts, the large amount of Poisson noise prevent those genes from exhibiting any DE at all, and DESeq performs independent filtering automatically to discard such low signals and to increase statistical power for the remaining gene candidates.

Analyse différentielle (2)

The asymptotic dispersion for highly expressed genes can be seen as a measurement of biological variability in the sense of a squared coefficient of variation: a dispersion value of 0.01 means that the gene's expression tends to differ by typically $\sqrt{0.01}$ = 10% between samples of the same treatment group. The R procedure `estimateDispersions` allows to compute (and visualize) dispersion estimates as a function of mean normalized counts.

Analyse différentielle (3)

The statistical test used to assess whether genes are differentially expressed between samples is a Wald test (`nbinomWaldTest`), with FDR correction for multiple testing. Benjamini–Hochberg's adjusted p-values can then be ranked to highlight the top genes. Usually, the statistical threshold is set at 0.1, and not 0.05 as in standard null hypothesis statistical testing.

The inspection of the distribution of (unadjusted) p-values is helpful to verify that the null distribution for the test statistic is viable. If the histogram does not exhibit an uniform pattern (e.g., U or hill shape), then it is likely that the $N(0, 1)$ null distribution is not appropriate. [2]

[2] See the fdrtool and locfdr packages for further strategies about controlling global or local FDR, and empirical null modeling allowing to estimate the variance of the null model (expected to be 1, per the $N(0, 1)$ hypothesis).

Références i

[1] Simon Anders et al. “Count-based differential expression analysis of RNA sequencing data using R and Bioconductor”. In: *Nature Protocols* 8.9 (2013), pp. 1765–1786.

[2] Ana Conesa et al. “A Survey of Best Practices for RNA-seq Data Analysis”. In: *Genome Biology* 17.13 (2016), pp. 1–19.

[3] Eija Korpelainen et al. *RNA-seq: Data Analysis A Practical Approach*. Taylor Francis/CRC Press, 2015.

[4] Michael I. Love, Wolfgang Huber, and Simon Anders. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. In: *Genome Biology* 15.550 (2014), pp. 1–21.

Références ii

[5] Craig R Yendrek, Elizabeth A. Ainsworth, and Jyothi Thimmapuram. “The Bench Scientist’s Guide to Statistical Analysis of RNA-Seq Data”. In: *BMC Research Notes* 5.506 (2012), pp. 1–10.