M2 Mycologie

Analyses RNA-Seq

Université Paris Cité

Principe des analyses RNA-Seq

- revues de bonnes pratiques : (Conesa et al. 2016; Yendrek, Ainsworth, et Thimmapuram 2012)
- outils statistiques : Bioconductor ; (Korpelainen et al. 2015; Anders et al. 2013)

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 (Love, Huber, et Anders 2014). 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)

Etape de normalisation : comparaison des facteurs de taille (rapport médian (ou moyenne tronquée avec le package edgeR) entre chaque échantillon et un échantillon virtuel de référence = médiane des valeurs pour chaque gène sur l'ensemble des échantillons). Ces rapports sont supposés tenir compte de la taille des librairies et être à peu égaux à 1. Si l'on divise chaque colonne de nombres de reads par le facteur de taille correspondant, on obtient le nombre de reads normalisé. On exprime celui-ci en unités par million pour l'interprétation.

Un graphique de type MA plot montre la moyenne en fonction de la différence moyenne de fold-change (en log), centré autour de 0. On s'attend à observer une plus grande variété des log ratios quand le nombre de reads est bas.

Analyse des données de comptage (3)

Une ACP ou une carte de contarste (”heatmap”) des résultats de la classification automatique des échantillons est utilisée pour vérifier la similarité entre les échantillons : les triplicats doivent être groupés ensemble et les échantillons provenant de conditions différentes doivent être éloignés les uns des autres. Généralement, on applique une transformation log régularisée sur les nombres de reads bruts pour minimiser l'impact des quelques gènes très variables, ce qui revient à donner un poids équivalent à tous les gènes. Pour les gènes avec un grand nombre de reads, cela équivaut à une transformation log2, alors que pour les gènes peu exprimés il s'agit plutôt de ramener les valeurs vers la valeur moyenne du gène.

Analyse différentielle

Le modèle statistique utilisé pour l'analyse différentielle repose sur une loi Binomiale négative. Contrairement à la loi de Poisson, cela permet de rendre compte de la surdispersion des valeurs (variance supérieure à la moyenne). La variance vaut $\mathbb{V}[NB(\mu, \alpha)] = \mu + \alpha\mu^2$, et la première étape de l'analyse consiste à estimer le paramètre de surdispersion (pour chaque gène, indépendemment de la condition).

Notons que pour les gènes avec un très faible nombre de reads, le bruit Poissonien annihile le moindre effet différentiel, et les outils d'analyse utilisent des filtres spéficiques pour supprimer ces gènes de l'analyse et augmenter la puissance statistique.

Analyse différentielle (2)

La dispersion asymptotique des gènes hautement exprimés peut être vue comme une mesure de la variabilité biologique (au sens d'un coefficient de variation au carré) : une valeur de dispersion de 0.1 signifie que l'expression du gène tend à différer par $\sqrt{0.01} = 10\%$ entre les échantillons de la même condition. La procédure R `estimateDispersions` permet de calculer et visualiser les valeurs estimées pour le paramètre de dispersion en fonction des valeurs de comptage normalisées.

Analyse différentielle (3)

Le test statistique utilisé pour évaluer si deux gènes sont différentiellement exprimés est un test de Wald (`nbinomWaldTest`), avec correction par FDR pour les tests multiples. Les p-valeurs ajustées de Benjamini–Hochberg peuvent être triées pour sousligner les "top gènes". Habituellement, le seuil est fixé à 0.1 et pas 0.05 comme dans le cadre des tests formels d'hypothèse.

La distribution des p-valeurs (non ajustées) est utile pour vérifier la distribution sous l'hypothèse nulle de la statistique de test. Si l'histogramme ne présente pas une allure uniforme (e.g., forme en U ou en V), alors il est vraisemblable que la distribution nulle $N(0, 1)$ null distribution n'est pas appropriée. [2]

[2] Voir les packages fdrtool et locfdr pour des stratégies alternatives de contrôle du FDR local ou global.

Références

Anders, Simon, Davis J. McCarthy, Yunshun Chen, Michal Okoniewski, Gordon K. Smyth, Wolfgang Huber, et Mark D. Robinson. 2013. « Count-based differential expression analysis of RNA sequencing data using R and Bioconductor ». *Nature Protocols* 8 (9): 1765‑86. https://doi.org/10.1038/nprot.2013.099.

Conesa, Ana, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, et Alejandra et al. Cervera. 2016. « A Survey of Best Practices for RNA-seq Data Analysis ». *Genome Biology* 17 (13): 1‑19. https://doi.org/10.1186/s13059-016-0881-8.

Korpelainen, Eija, Jarno Tuimala, Panu Somervuo, Mikael Huss, et Garry Wong. 2015. *RNA-seq: Data Analysis A Practical Approach*. Taylor & Francis/CRC Press. https://doi.org/10.1201/b17457.

Love, Michael I., Wolfgang Huber, et Simon Anders. 2014. « Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 ». *Genome Biology* 15 (550): 1‑21. https://doi.org/10.1186/s13059-014-0550-8.

Yendrek, Craig R, Elizabeth A. Ainsworth, et Jyothi Thimmapuram. 2012. « The Bench Scientist's Guide to Statistical Analysis of RNA-Seq Data ». *BMC Research Notes* 5 (506): 1‑10. https://doi.org/10.1201/b16589-3.