Jeremy Ferraro - 14/10/25 - Université Paris-Saclay, CNRS, INRAE, Université Evry, Université Paris Cité, Institute of Plant Sciences Paris-Saclay (IPS2), Gif sur Yvette 91190, France

Impact du nombre de réplicats sur les analyses omiques

Les approches dites *omiques* regroupent l'ensemble des disciplines visant à mesurer globalement un niveau moléculaire donné — gènes, ARN, protéines ou métabolites — afin de décrire l'organisation et les interactions des systèmes biologiques. Si ces technologies offrent une vision étendue de l'activité cellulaire, leur mise en œuvre soulève d'importants défis analytiques. Le nombre souvent limité de réplicats biologiques, comparé au volume de variables mesurées, diminue la puissance statistique, autrement dit, la probabilité de détecter une molécule réellement différentielle entre deux conditions ([1], [2]). Dans ce contexte, les études omiques cherchent fréquemment à identifier les molécules spécifiquement associées à une condition donnée. Ce sous-ensemble d'entités spécifiques constitue souvent la base de l'interprétation biologique.

Ainsi, ce travail vise à évaluer l'impact du nombre de réplicats sur l'interprétation biologique des analyses omiques. D'une part, il s'agit de déterminer comment le manque de réplicats affecte la reproductibilité et l'interprétation des approches mono-omiques — transcriptomique et protéomique — et de tester si les analyses d'enrichissement fonctionnel peuvent compenser, au moins partiellement, la perte d'information liée à la faible puissance. D'autre part, l'objectif est également d'étudier comment ces limitations influencent la cohérence entre niveaux moléculaires lors de l'intégration multi-omique.

Pour répondre à ces questions, nous avons exploité un jeu de données constitué de 22 réplicats biologiques d'*Arabidopsis thaliana* soumis à un double stress chaleur et CO_2, avec des mesures transcriptomiques et protéomiques. Ensuite, nous avons mis en place une stratégie de sous-échantillonnage, consistant à répéter aléatoirement les analyses avec un sous-ensemble restreint de réplicats. Cette approche permet ainsi de simuler des designs expérimentaux plus classiques (3 réplicats) et de comparer les performances à celles obtenues avec l'ensemble des données.

Notre analyse met en évidence une faible reproductibilité des résultats lorsque le nombre de réplicats est limité: les listes de gènes différentiellement exprimés varient fortement d'un sous-échantillonnage à l'autre. Face au manque de puissance et à la forte variabilité observée entre sous-échantillonnages, nous avons testé la spécificité des réponses en analysant des diagrammes de Venn, couramment utilisés pour visualiser les gènes différentiellement exprimés entre conditions simples et combinées.

Nous observons qu'avec trois réplicats sur vingt-deux, des centaines de gènes apparaissent comme spécifiques d'une condition combinée (CO_2 + température). Pourtant, lorsque ces résultats sont comparés à ceux obtenus avec l'ensemble des réplicats, une grande partie de ces gènes est également détectée comme différentielle dans les conditions simples. Autrement dit, de nombreux gènes interprétés comme "spécifiques" d'une combinaison de stress ne le sont qu'en apparence: leur statut résulte du manque de puissance, et non d'une régulation réellement exclusive.

Ce phénomène ne se limite pas à cette expérience. Lorsque la puissance est faible, la probabilité de détecter simultanément un même gène comme différentiel dans plusieurs conditions chute mécaniquement, ce qui crée artificiellement des "spécificités" de condition. Ainsi, même dans un scénario où toutes les régulations seraient réellement partagées, des diagrammes de Venn construits à partir de données avec peu de réplicats conduiraient inévitablement à une inflation de faux positifs et à une fausse impression de spécificité.

Les analyses d'enrichissement ne corrigent pas ces biais: lorsque le nombre de réplicats est limité, elles reflètent la même incertitude que les listes de gènes différentiels sur lesquelles elles reposent, les voies ou catégories identifiées comme significativement enrichies varient d'un sous-échantillonnage à l'autre. Les premières observations en protéomique vont dans le même sens : les effets du manque de puissance et de reproductibilité y sont tout aussi perceptibles. Les faibles concordances souvent rapportées entre données transcriptomiques et protéomiques ne traduisent donc pas nécessairement des divergences biologiques réelles, mais résulteraient en grande partie de la sensibilité limitée des analyses elles-mêmes.

Ce constat appelle à une réflexion sur le design expérimental des expériences omiques. L'idéal reste évidemment d'augmenter le nombre de réplicats biologiques, mais cette approche se heurte à des contraintes expérimentales bien réelles — coûts, disponibilité des infrastructures, ou encore complexité technique des manipulations.

Une alternative consiste à reformuler la question biologique en se concentrant non pas sur les gènes "uniquement spécifiques" à une condition combinée, mais sur les interactions entre facteurs, en utilisant des modèles statistiques comportant un terme d'interaction. Cette approche permet de contrôler le taux de faux positifs, mais elle requiert un plan factoriel complet, incluant toutes les combinaisons possibles de stress. Lorsque le nombre de conditions devient trop important, un plan factoriel complet n'est plus envisageable; les plans factoriels fractionnaires offrent alors un compromis réaliste, dans lesquels certaines combinaisons sont volontairement omises. Cette stratégie permet de réduire considérablement la charge expérimentale, tout en conservant la possibilité d'estimer les effets principaux et certaines interactions. Pour les analyses visant à identifier des différences spécifiques, il est possible de recourir à des approches statistiques plus adaptées, telles que les contrôles post hoc du taux de faux positifs ou des approches de tests d'hypothèses composites, afin de mieux encadrer l'interprétation des résultats.

[1] N. J. Schurch *et al.*, « How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? », *RNA*, vol. 22, nº 6, p. 839-851, juin 2016, doi: 10.1261/rna.053959.115.

[2] S. Lamarre *et al.*, « Optimization of an RNA-Seq Differential Gene Expression Analysis Depending on Biological Replicate Number and Library Size », *Front. Plant Sci.*, vol. 9, 2018, Consulté le: 25 juillet 2023. [En ligne]. Disponible sur: https://www.frontiersin.org/articles/10.3389/fpls.2018.00108

Ce travail a été effectué en collaboration avec **Jeremy Ferraro, Virginie Noël, Axel de Zelicourt, Michael Hodges, Elodie Gilbault, Olivier Loudet, José Caius, Alexandra Launay-Avon, Stéphanie Pateyron, Christine Paysant Le Roux, Marie-Laure Martin, Benoît Castandet, Etienne Delannoy, Guillem Rigaill.** Une partie de ces travaux est en cours de publication.