Inférence post-clustering pour l'identification des variables responsables de la séparation de paires de clusters

Benjamin Hivert[1,2,3,*], Denis Agniel[4,5], Rodolphe Thiébaut[1,2,3,6], Boris Hejblum[1,2,3]

[1] Univ. Bordeaux, Inserm Bordeaux Population Health Research Center, SISTM team, UMR 1219, Bordeaux F33076, France
[2] INRIA Bordeaux Sud Ouest, SISTM team Talence F-33400, France
[3] Vaccine Research Institute, VRI, Hôpital Henri Mondor, Créteil F-94000, France
[4] Rand Corporation, Santa Monica, CA 90401, USA
[5] Harvard Medical School, Boston, MA 02115, USA
[6] CHU Pellegrin, Groupe Hospitalier Pellegrin, Bordeaux F-33076, France
[*] benjamin.hivert@u-bordeaux.fr

L'analyse de données d'expression génique RNA-seq s'organise souvent autour de deux étapes successives : i) un *clustering* pour grouper les unités d'observation (patients, cellules, …) en sous-groupes homogènes et séparés ; puis ii) une analyse différentielle à l'aide de tests d'hypothèses pour identifier quels gènes, c'est-à-dire quelles variables, séparent chacun des sous-groupes. Il est cependant possible que plusieurs sous-groupes construits en i) ne contiennent en réalité que des observations provenant d'une même population homogène. Le *clustering* va alors artificiellement créer des différences entre les *sous-groupes,* qui seront ensuite identifiées en ii) de façon erronée. Or ces différences induites par le *clustering* ne viennent pas d'un processus biologique qui sépareraient nos observations, mais simplement de cette double utilisation des données générant une inflation de l'erreur de type I. Il est important de proposer des méthodes d'inférence permettant de tenir compte de l'étape de clustering afin de pouvoir répondre à la question suivante : est-ce que la différence observée entre deux sous-groupes selon une variable est uniquement due au fait qu'une méthode de clustering a été appliquée au préalable sur les données, ou si au contraire cette différence existe indépendamment du *clustering*.
Ce problème d'inférence post-clustering fait l'objet de développements récents [1,2]. Nous proposons deux méthodes d'inférence permettant de tenir compte de l'étape de clustering dans l'analyse différentielle : a) nous étendons le travail de Gao et al. [2] au cas uni-variable qui s'appuie sur le concept d'inférence sélective [3] où l'on conditionne sur le *clustering* dans le test statistique ; b) avec une définition plus restrictive d'un sous-groupe (utilisant les notions d'unimodalité et de multimodalité pour caractériser respectivement l'homogénéité au sein d'un sous-groupe et la séparation entre deux sous-groupes) nous étudions la séparabilité de deux *clusters* selon une variable d'intérêt par un test de multimodalité.
Les deux méthodes proposées conduisent à des p-valeurs valides sous l'hypothèse nulle (absence de différence entre sous-groupes selon la variable d'intérêt indépendamment du *clustering*). Cependant, ces méthodes sont très dépendantes de la définition de sous-groupe qu'elles utilisent. De plus, leurs performances en grande dimension sont nettement inférieures à celles en petite dimension, notamment car les nombreux groupes de variables corrélées peuvent induire des sous-groupes homogènes et donc interprétables, qui ne sont néanmoins jamais suffisamment séparés en uni-variable pour respecter les définitions de sous-groupes utilisées dans chacune des deux approches.

Références :

[1] Zhang, J. M., Kamath, G. M., & David, N. T. (2019). Valid post-clustering differential analysis for single-cell RNA-Seq. *Cell systems*, *9*(4), 383-392.
[2] Gao, L. L., Bien, J., & Witten, D. (2020). Selective Inference for Hierarchical Clustering. *arXiv preprint arXiv:2012.02936*.
[3] Tibshirani, R. J., Taylor, J., Lockhart, R., & Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, *111*(514), 600-620.