Tests de fonction de répartition cumulative conditionnelle pour l'analyse d'ensembles de gènes de données RNA-seq en cellule unique

Sara Fallet[1,2], Denis Agniel[3], Rodolphe Thiébaut[1,2,4] & Boris Hejblum[1,2]

[1] *Univ. Bordeaux, INSERM, INRIA, Bordeaux Population Health, SISTM team, U1219, F-33000 Bordeaux, France*
[2] *Vaccine Research Institute, VRI, Hôpital Henri Mondor, Créteil F-94000, France*
[3] *Rand Corporation, Santa Monica, CA 90401, USA*
[4] *CHU Pellegrin, Groupe Hospitalier Pellegrin, Bordeaux F-33076, France*

sara.fallet@u-bordeaux.fr, dagniel@rand.org
rodolphe.thiebaut@u-bordeaux.fr, boris.hejblum@u-bordeaux.fr

Résumé. La technologie de séquençage d'ARN en cellule unique (scRNA-seq) mesure l'expression génique dans des centaines, voire des milliers de cellules à partir d'un seul échantillon biologique, permettant d'étudier les mécanismes moléculaires à l'échelle de la cellule. En immunologie, cette technologie est de plus en plus utilisée pour étudier la réponse immunitaire lors d'une infection (ou vaccination) tout en tenant compte de l'hétérogénéité cellulaire dans le sang. L'analyse de l'expression différentielle identifie les gènes dont l'expression change selon les différentes conditions d'étude. Cependant, les méthodes d'analyse différentielle manquent de puissance statistique et de stabilité, notamment en raison de la nature très dynamique de l'expression génique, de l'hétérogénéité de l'état cellulaire et des limitations technologiques telles que la profondeur de séquençage. En s'intéressant plutôt à des ensembles de gènes associés à des fonctions immunitaires spécifiques, définis à partir de connaissances biologiques *a priori*, on améliore la puissance statistique et la stabilité de l'analyse tout en facilitant l'interprétation biologique des résultats. Nous présentons ici une nouvelle méthode d'analyse différentielle par groupes de gènes adaptée aux données scRNA-seq. Cette méthode repose sur une estimation suivie d'un test de la fonction de répartition conditionnelle de l'expression des gènes au sein d'un groupe. Cette nouvelle méthode s'affranchit ainsi du besoin de faire une hypothèse distributionnelle (délicate pour les données scRNA-seq). Elle est également capable d'analyser des plans expérimentaux complexes, testant l'association de chaque ensemble de gènes avec une ou plusieurs variables d'intérêt (qu'elles soient continues ou discrètes), tout en ajustant éventuellement sur d'autres covariables, dépassant ainsi le cadre usuel de la comparaison simple entre deux groupes. Nous appliquons cette nouvelle méthodologie à deux jeux de données réelles de scRNA-seq étudiant la réponse immunitaire à l'infection par le SARS-CoV-2 chez l'homme, avec respectivement 84 140 cellules T CD8+ provenant de 38 patients et 1 191 463 cellules mononucléaires du sang périphérique provenant de 222 donneurs.

Mots-clés. séquençage d'ARN en cellule unique, analyse par groupes de gènes, analyse d'expression différentielle, test d'indépendance conditionnel, fonction de répartition cumulative conditionnelle

Abstract. Single-cell RNA-seq (scRNA-seq) technology measures gene expression in hundreds or even thousands of cells from a single biological sample, allowing to study molecular mechanisms at the single-cell resolution. In immunology, this technology is increasingly used to disentangle the complex immune response to infection (or vaccination) while accounting for cellular heterogeneity in the blood. Differential Expression Analysis (DEA) allows to identify which genes are differentially expressed across different conditions, cell types, timings or exposures. However, DEAs often encounter challenges related to statistical power and stability, notably due to the dynamic nature of gene expression and cellular state heterogeneity. Investigating instead gene sets associated with specific immune functions, derived from prior biological knowledge, can enhance the statistical power and stability of the analysis while facilitating the biological interpretation of results.
We introduce a novel gene set analysis method tailored for scRNA-seq data. This method relies on the estimation and testing of conditional distribution functions, eliminating the need for distributional assumptions. This new method is suitable for complex experimental designs, testing the association of each gene set with one or multiple variables of interest (whether continuous or discrete), while potentially adjusting for additional covariates. We apply this new methodology to two single-cell RNA-seq real dataset investigating the immune response to SARS-CoV-2 infection in humans, with respectively 84,140 virus-reactive CD8+ T cells from 38 patients; and 1,191,463 peripheral blood mononuclear cells from 222 donors.

Keywords. single cell RNA-seq, gene set analysis, differential expression analysis, conditional independence test, conditional cumulative distribution function

1 Introduction

Le séquençage d'ARN en cellule unique permet une analyse plus fine de l'expression génique, permettant la mesure de l'expression génique de chaque cellule individuellement. Comparé au séquençage de l'ARN en masse (*bulk*), qui traite les cellules comme un mélange homogène sans les distinguer, le scRNA-seq offre une résolution plus précise en tenant compte des différences entre les types cellulaires et leurs états. Les données à cellule unique présentent des caractéristiques différentes des données de séquençage d'ARN en masse qui nécessitent une considération spéciale pour le développement d'outils DEA. Notamment, les données scRNA-seq affichent de grandes proportions de zéros, en raison de processus biologiques ou de limitations techniques. Ces données sont également multi-échelles puisqu'elles ont une résolution au niveau des cellules, des échantillons et des conditions. Comme l'analyse des données scRNA-seq présente des particularités uniques, de nouvelles méthodes statistiques sont nécessaires. Les méthodes existantes font des hypothèses fortes sur la distribution des données qui sont très difficiles à vérifier en pratique. Il est donc important de développer des méthodes générales et flexibles pour analyser les données scRNA-seq qui ne nécessitent pas de fortes hypothèses paramétriques.

L'analyse différentielle de l'expression génique permet d'identifier les gènes dont l'expression change entre différentes conditions d'étude (par exemple types cellulaires, temps ou expositions). Un gène est ainsi appelé différentiellement exprimé si son expression est significativement associée aux variations d'un facteur d'intérêt. La plupart des méthodes

existantes pour l'analyse différentielle de données scRNA-seq permettent seulement la comparaison entre deux groupes. En pratique, on peut s'intéresser à des schémas expérimentaux plus complexes, qui nécessitent la comparaison de plus de deux groupes – comme par exemple plusieurs niveau de sévérités d'une maladie.

Nous proposons une nouvelle méthode basée sur les travaux de Gauthier et al. (2021), `citcdf gsa`, qui est non paramétrique et permet de comparer plus de 2 conditions. Elle s'appuie sur un test d'indépendance conditionnel entre l'expression d'un gène Y et une variable d'intérêt X, ajusté sur des covariables Z. Ce test estime la fonction de répartition conditionnelle (à l'aide de multiples régressions logistiques) afin d'évaluer la potentiel différence selon le conditionnement sur X. Les performances de notre méthode ont été étudiées par des simulations numérique ainsi que pour l'analyse de deux jeux de données réelles scRNA-seq étudiant chacun la réponse immunitaire à l'infection par le SARS-CoV-2.

2 Méthode

2.1 `citcdf` : analyse gène par gène (Gauthier et al., 2021)

On effectue un test d'indépendance conditionnel entre l'expression d'un gene Y et un facteur d'intérêt X, sachant des covariables Z :

$$H_0 : Y \perp X \mid Z,$$

où X et Z peuvent être respectivement le statut d'une maladie ou le bras d'un vaccin, et le type d'une cellule ou l'âge du patient par exemple.

Notre méthode d'analyse d'expression différentielle se base sur l'estimation de fonction de répartition conditionnelle. Ainsi l'hypothèse nulle de notre test peut-être reformulée comme :

$$H_0 : F_{Y|X,Z}(y, x, z) = F_{Y|Z}(y, z),$$

où $F_{Y|X,Z}(y, x, z)$ est la fonction de répartition conditionnelle de Y sachant X et Z, et $F_{Y|Z}(y, z)$ est la fonction de répartition « marginale » (au sens indépendemment de X) de Y (mais toujours conditionnellement à Z). On teste l'égalité entre ces deux fonctions car, si un groupe de facteur est associé à l'expression d'un gène, alors la $F_{Y|X,Z}$ sera significativement différente de $F_{Y|Z}$.

La fonction de répartition conditionnelle peut s'exprimer sous la forme d'une probabilité :

$$F_{Y|X,Z}(y, x, z) = \mathbb{P}(Y \leq y | X = x, Z = z).$$

Pour chaque gène, on peut ainsi choisir une séquence de p seuils réguliers et ordonnés ω_j où $j = 1, \ldots, p$. Ainsi $F_{Y|X,Z}$ peut s'écrire comme l'espérance conditionnelle d'un variable binaire :

$$F_{Y|X,Z}(y, x, z) = \mathbb{E}(\mathbb{1}_{\{Y_i \leq \omega_j\}} | X_i = x, Z_i = z) \tag{1}$$

On peut alors estimer (1) par un séquence de p régressions logistiques :

$$g(\mathbb{E}[\mathbb{1}_{\{Y_i \leq \omega_j\}} | X_i, Z_i]) = \beta_{0j} + \boldsymbol{\beta}_{1j}\mathbf{X_i} + \boldsymbol{\beta}_{2j}\mathbf{Z_i},$$

où $i = 1, \ldots, n$ indexe les observations de chaque cellule, $\boldsymbol{\beta}_{1j} = (\beta_{1j1}, \ldots, \beta_{1js_1})$ quantifie l'influence de X_i sur $\mathbb{P}(Y_i \leq \omega_j)$ et $\boldsymbol{\beta}_{2j} = (\beta_{2j1}, \ldots, \beta_{2js_2})$ quantifie l'influence de Z_i sur $\mathbb{P}(Y_i \leq \omega_j)$. s_1 et s_2 désignent respectivement le nombre total de conditions et de variables considéré.

Si le facteur d'intérêt X n'a aucun lien avec l'expression du gène Y, $\boldsymbol{\beta}_{1j}$ sera égal à 0. De ce fait l'hypothèse nulle de notre test devient :

$$H_0 : \boldsymbol{\beta}_1 = 0,$$

où $\boldsymbol{\beta}_1 = (\boldsymbol{\beta}_{11}, \ldots, \boldsymbol{\beta}_{1j}, \ldots, \boldsymbol{\beta}_{1p})$ est la matrice des β_1 pour chaque seuil d'expression choisi et pour chaque condition testée.

La statistique de test du score associée peut s'estimer comme $\widehat{D}_n = n \sum_{j=1}^{p} \sum_{k=1}^{s_1} \widehat{\beta}_{1jk}^2$, ainsiq que sa distribution asymptotique $\widehat{D}_n \underset{n \to +\infty}{\longrightarrow} \sum_{q=1}^{ps_1} \widehat{a}_q \chi_1^2$ avec les $\widehat{\beta}_{1jk}$ qui sont approchés grâce à la méthode des moindres carrés ordinaires (afin d'accélérer les calculs), tandis que les $\widehat{a}_q$ sont les valeurs propres de la matrice de variance covariance des β_{1jk} notée Σ. Enfin, on calcule les p-valeurs en comparant la statistique de test observée $\widehat{D}_n$ avec sa distribution asymptotique.

2.2 `citcdf gsa` : analyse par groupe de gènes

Dans le cadre de l'analyse par groupe de gènes, on estime cette fois (1) avec le modèle :

$$g(\mathbb{E}[\mathbb{1}_{\{Y_i \leq \omega_j\}} | X_i, Z_i]) = \beta_{0jl} + \boldsymbol{\beta}_{1jl} \mathbf{X_i} + \boldsymbol{\beta}_{2jl} \mathbf{Z_i},$$

où $l = 1, \ldots, m$ désigne les différents gènes du groupe de gènes.

Notre hypothèse nulle devient naturellement $H_0 : \boldsymbol{\beta}_1 = 0$, avec $\boldsymbol{\beta}_1 = (\boldsymbol{\beta}_{111}, \ldots, \boldsymbol{\beta}_{1p1}, \ldots, \boldsymbol{\beta}_{11m}, \ldots, \boldsymbol{\beta}_{1pm})$ la matrice des β_{1jlk} pour chaque seuils, de chaque gène du groupe de gène. Ainsi la statistique de test et sa distribution asymptotique deviennent :

$$\widehat{S}_n = n \sum_{j=1}^{p} \sum_{k=1}^{s_1} \sum_{l=1}^{m} \widehat{\beta}_{1jkl}^2 \qquad \widehat{S}_n \underset{n \to +\infty}{\longrightarrow} \sum_{j=1}^{ps_1 m} \widehat{a}_j \chi_1^2$$

Lors du calcul de la matrice de variance covariance $\widehat{\Sigma}$, qui permet d'estimer les valeurs propres $\widehat{a}_j$, on a des termes croisés supplémentaires entre les $\widehat{\boldsymbol{\beta}}_{1jkl}$ de chaque gène du groupe set.

3 Résultats

3.1 Simulations numériques

Afin d'évaluer les performances du test proposé dans `citcdf gsa`, nous avons généré des simulations numériques gaussiennes, paramétrées avec $n = 100$ cellules et $r = 200$ gènes. Des facteurs d'intérêt X_1 et X_2, et des covariables Z_1 et Z_2 ont été générés à partir de distributions normales, et finalement la matrice Y d'expression des gènes a également été générées grâce à l'addition d'un bruit gaussien. Les résultats testant des groupes de deux gènes indépendant ont été agrégés sur 1500 simulations, et sont représentés en Figure 1. On constate un controle adéquate de l'erreur de Type-I sous H_0, accompagné d'une puissance adéquate sous H_1.

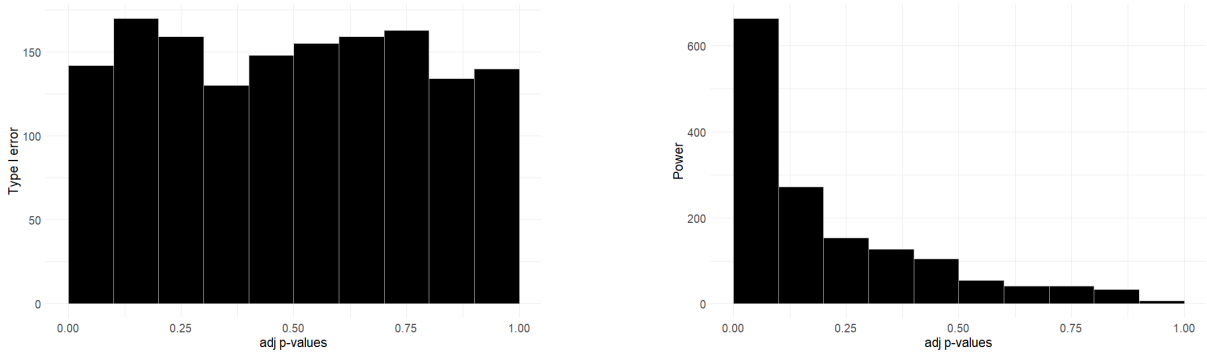

FIGURE 1 : Performance du test en simulation. À gauche : test sous l'hypothèse nulle H_0. À droite : test sous l'hypothèse alternative H_1

3.2 Applications à des données scRNA-seq réelles

Kusnadi et al. (2021) ont mesuré l'expression de 13 816 gènes au sein de 84 140 cellules T CD8 positives, réactives au virus SARS-CoV-2, chez 39 patients atteints de la COVID-19 répartis en trois groupes : 17 patients non hospitalisés, 13 hospitalisés et 9 hospitalisés en unité de soins intensifs. Kusnadi et al. (2021) ont mis en évidence l'importance de 6 groupes de gènes *Exhaustion Consensus*, *Genes upregulated in cytotoxicity*, *Type I and II Inferon signaling genes*, *Viral response*, *Hallmark Glycolysis* et *Genes upregulated in cell cycle* contenant entre 11 et 226 gènes. `citcdf gsa` identifie l'ensemble de ces 6 groupes de gènes comme significatif avec un taux de fausse découverte inférieur à 5%. La Figure 2 représente l'expression des gènes du groupe de gènes annotés *Genes upregulated in cytotoxicity* regroupé selon les 3 sévérités cliniques.

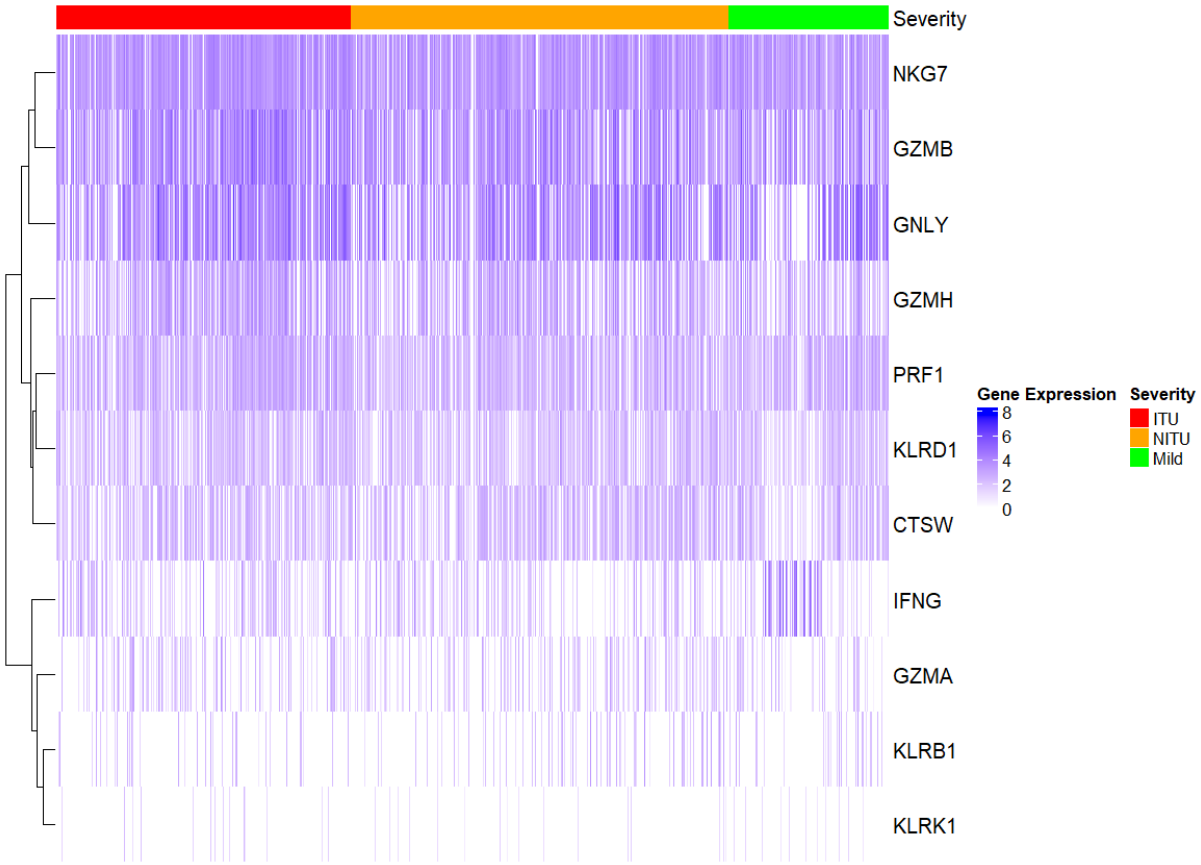

FIGURE 2 : Heatmap de l'expression du groupe de gènes *Genes upregulated in cytotoxicity*

4 Discussion

`citcdf gsa` est une méthode d'analyse d'expression différentielle par groupe de gène applicable aux données scRNA-seq. Elle réalise un test d'indépendance conditionnel qui se base sur l'estimation de la fonction de répartition conditionnelle grâce à de multiple régressions logistiques. Elle permet de prendre en compte des schémas expérimentaux complexes, notamment en permettant d'étudier plus de 2 conditions expérimentales ou en permettant d'ajuster sur des covariables. Étant une méthode non-paramétrique, `citcdf gsa` peut suivre n'importe quelle méthode de normalisation des données de séquençage. `citcdf gsa` présente de bonnes performances en simulation avec un contrôle efficace de l'erreur de type I et une puissance statistique raisonnable.

Nous avons implémenté cette méthode dans le package R `citcdf`, branche `gsa` pour l'analyse par groupe de gènes, disponible sur *GitHub* : `https://github.com/sistm/citcdf/tree/gsa`.

Références

Gauthier, M., Agniel, D., Thiébaut, R., and Hejblum, B. P. (2021). Distribution-free complex hypothesis testing for single-cell RNA-seq differential expression analysis. Preprint, Bioinformatics.

Kusnadi, A., Ramírez-Suástegui, C., Fajardo, V., Chee, S. J., Meckiff, B. J., Simon, H., Pelosi, E., Seumois, G., Ay, F., Vijayanand, P., and Ottensmeier, C. H. (2021). Severely ill patients with COVID-19 display impaired exhaustion features in SARS-CoV-2–reactive CD8 $^{+}$ T cells. *Science Immunology*, 6(55):eabe4782.