# IIDEA : Interactive Inference for Differential Expression Analyses

Nicolas Enjalbert Courrech* Pierre Neuvial†

**Résumé**

Nous avons développé l'application shiny IIDEA (Interactive Inference for Differential Expression Analyses) pour l'analyse différentielle d'expression de gènes. Celle-ci exploite les développements récents en statistique dans le domaine de l'inférence post hoc, qui fournissent des garanties statistiques plus interprétables que les méthodes classiques comme le contrôle du taux de fausses découvertes (FDR). En particulier, l'utilisateur d'IIDEA peut de façon interactive sélectionner des gènes d'intérêt à l'aide d'un volcano plot ou d'annotations biologiques, et obtenir une garantie sur le nombre ou la proportion de gènes différentiellement exprimés dans cette sélection. IIDEA est déployée sur le lien suivant : https://shiny-iidea-sanssouci.apps.math.cnrs.fr/ .

**Mots-clefs** : shiny – dataviz – expression différentielle – volcano plot – analyse d'enrichissement – inférence post hoc

## Introduction

Les études d'expression différentielle visent à identifier les gènes dont l'expression moyenne diffère significativement entre deux groupes d'individus. L'analyse statistique standard de ces données consiste à réaliser un test statistique par gène, puis une correction de test multiple de type False Discovery Rate (FDR, proportion de faux positifs attendus). En sortie l'utilisateur obtient une liste de gènes dont le FDR est contrôlé. Un inconvénient majeur de ce type d'analyse est qu'elle ne fournit une garantie statistique que sur cette liste, et pas par exemple à des sous-ensembles ou sur-ensembles, voir Ebrahimpoor and Goeman (2021). La notion d'*inférence post hoc* introduite par Goeman and Solari (2011) répond à ce besoin en fournissant des garanties statistiques sur le nombre ou la proportion d'erreurs dans n'importe quelle liste de gènes définie par l'utilisateur.

Afin de rendre ces méthodes utilisables pour des non programmeurs, et de tirer parti des possibilités d'interaction qu'elles offrent, nous avons créé une interface Shiny (voir Chang et al. (2021)) afin de rendre la sélection des gènes réactive, fonctionnelle et facile d'utilisation. L'application IIDEA est déployée à l'adresse: https://shiny-iidea-sanssouci.apps.math.cnrs.fr/. Elle repose sur les méthodes post hoc développées par Blanchard, Neuvial, and Roquain (2020) et implémentées par Neuvial et al. (2021) dans le package `sansSouci`.

## Aperçu des fonctionnalités

Les principales fonctionnalités sont illustrées par la Figure 1: paramètres utilisateur (cadre bleu pointillé), volcano plot (cadre vert en trait pointillé) et table des bornes post hoc (cadre rouge). Un volcano plot est un graphique permettant une lecture rapide des gènes distinguant deux populations. Avec une mesure de fold-change (taille d'effet) en abscisse et une mesure de significativité en ordonnée, en échelle $-log$, les points situés dans les coins hauts droit et gauche représentent les gènes qui permettent de faire la distinction entre les deux populations. L'application permet de faire une sélection en coin grâce à des seuils sur le fold-change et la significativité déplaçables par glisser-déposer.

*Institut de Mathématiques de Toulouse, nicolas.enjalbert-courrech@math.univ-tlse3.fr
†Institut de Mathématiques de Toulouse, pierre.neuvial@math.cnrs.fr

Les méthodes post hoc de Blanchard, Neuvial, and Roquain (2020) sont alors appliquées à chaque sélection, et résumées dans le tableau encadré en rouge. L'utilisateur a également la possibilité de sélectionner des ensembles de gènes d'intérêt provenant de bases de données d'annotation comme Gene Ontology. Les bornes post hoc sont calculées pour chaque ensemble de gènes.

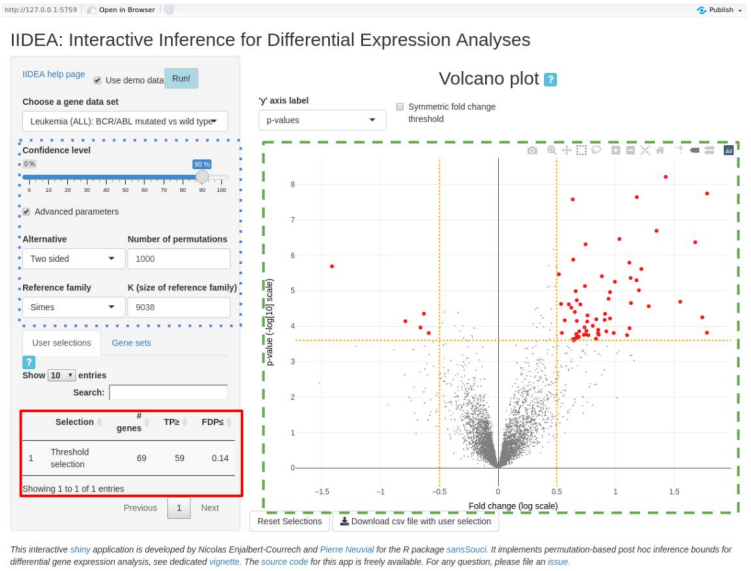


Figure 1: Capture d'écran de l'application IIDEA

Afin d'augmenter la réactivité de l'application, nous avons utilisé un graphique plotly multi-couche (voir Sievert (2020)) pour optimiser le temps d'affichage du volcano plot. Avec cette technique, l'ensemble des points et les points sélectionnés ne sont pas sur la même couche. Ainsi la modification de la sélection ne va pas réimprimer le graphique entier mais uniquement la couche concernée.

## Références

Blanchard, Gilles, Pierre Neuvial, and Etienne Roquain. 2020. "Post Hoc Confidence Bounds on False Positives Using Reference Families." *Annals of Statistics* 48 (3): 1281–1303. https://projecteuclid.org/euclid.aos/1594972818.

Chang, Winston, Joe Cheng, JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert, and Barbara Borges. 2021. *Shiny: Web Application Framework for R.* https://CRAN.R-project.org/package=shiny.

Ebrahimpoor, Mitra, and Jelle J Goeman. 2021. "Inflated False Discovery Rate Due to Volcano Plots: Problem and Solutions." *Briefings in Bioinformatics.*

Goeman, Jelle J, and Aldo Solari. 2011. "Multiple Testing for Exploratory Research." *Statistical Science* 26 (4): 584–97.

Neuvial, Pierre, Guillermo Durand, Nicolas Enjalbert-Courrech, and Marie Perrot-Dockès. 2021. *SansSouci: Post Hoc Multiple Testing Inference.* https://pneuvial.github.io/sanssouci.

Sievert, Carson. 2020. *Interactive Web-Based Data Visualization with R, Plotly, and Shiny.* Chapman; Hall/CRC. https://plotly-r.com.