**PROBLÉMATIQUE (donnée par le prof) :**

Expliquez quelles sont les faiblesses (sur le plan statistique) des stratégies classiques pour l'analyse différentielle (contrôle du FDR) et si et pourquoi les approches proposées permettent de surmonter ces faiblesses.

**INTRODUCTION :**

La procédure la plus populaire de contrôle du False Discovery Rate (FDR) (erreur de type I) est la procédure de Benjamini-Hochberg utilisant seulement les p-valeurs ajustées. Une autre procédure est le vulcano plot. Cette dernière utilise à la fois les p-valeurs ajustées et l’effet de taille. Cependant, la procédure BH ne garantit pas le contrôle des erreurs sur les sous-ensembles et donc peut être important.

**OBJECTIFS :**

Problématique du prof ?

**MÉTHODES (A compléter avec le cours…) :**

* Benjamini-Hochberg : corrige les p-valeurs et maintient un niveau FDR à 0,05 (en général). La méthode BH seule trouve trop de gènes significatifs donc pour remédier à cela on fait une procédure de “double filtrage".
* Procédure de "double filtrage" (vulcano plot) : 1) les résultats sont filtrés en fonction de la signification; 2) filtrage grande taille de l’effet.

Le graphique en forme de volcan est un diagramme de dispersion de la signification statistique (valeurs -log10p sur l'axe des ordonnées) par rapport à l'ampleur de l'effet (taille de l'effet estimé ou variation en plis sur l'axe des abscisses). Les variables les plus significatives sont en haut dans les coins.

Le problème de cette procédure est qu'elle ne tient pas compte du fait que la garantie du contrôle FDR ne s'applique qu'à l'ensemble complet corrigé par BH, et pas nécessairement au sous-ensemble filtré. FDR = nb faux positif (FDP)/ nb tot de positif. → Pour conserver le contrôle du FDR il faut que FDP et nb tot de positif soit réduit en même temps, or ici ce n’est pas forcément le cas.

**RÉSULTATS :**

Problème double filtrage :

* Simulation : Figure 2 : Plus il y a surreprésentation des grandes valeurs de sigma parmi les gènes non nuls (lambda) plus le nombre de gènes sélectionnés augmente et moins la FDP est bonne.

Figure 3 : Plus la taille d’échantillon augmente, moins la FDP est bonne. De plus, lorsque la taille d’effet augmente la FDP devient meilleure.

Le FDR gonfle lorsqu' il y a une grande taille d’échantillon et lorsque la distribution des écart-types des différentes.

De plus, la FDR devient plus faible dans certains scénarios si les gènes sont corrélés, mais le problème ne disparaît pas.

* Étude réelle : 49 individus et 58037 gênes. Méthode de rééchantillonnage avec 6 ind pour l’échantillon test et 37 ind pour l’échantillon de validation. Si parmi les 37 il y a des gènes significatifs pour le BH (valeur P ajustée pour le BH <0,05) alors ils sont considérés comme différentiellement exprimés. Si DE dans validation et test → VP; si DE dans validation et non DE dans test (ou inversement) → FP; si non DE dans val et test → VN.

17 % des gènes sélectionnés par le vulcano plot dans test n'étaient pas significatifs dans la validation. Le filtrage des résultats peut conduire à une inflation du FDP et qu'une taille d'effet importante n'implique pas nécessairement que le résultat est un vrai positif.

Solutions :

* Focused BH (à compléter) : est une variante de la procédure BH classique qui garantit le contrôle de la FDR sur un sous-ensemble de gènes sélectionnés de manière prédéfinie.
* Closed testing (pas tout compris…): contrôle la FDP sur tous les sous-ensembles possibles. Cette méthode fournit une estimation et un IC de la FDP. Elle permet un contrôle de la FDP, quelle que soit la manière dont les gènes ont été sélectionnés.
* Résultats :
  + sim1 : le lambda a été contrôlé mais l’effet de taille joue encore un rôle important quand il est élevé (figure 6). On a de plus que l’estimateur par focused BH est toujours légèrement supérieur à l’estimateur de closed testing. Enfin, les 2 méthodes sous-estiment le nombre de vraies valeurs.
  + sim2 (la valeur de τ ne dépend pas des données) : pour les 2 méthodes, l'estimation de la FDP se rapproche de la vérité lorsque le seuil P diminue. Les 2 méthodes surestiment la vrai FDP. (figure 7)
  + sim3 (la valeur de τ dépend des données) : même idée que la sim2. Les 2 méthodes surestiment la vrai FDP. (figure 8)
  + étude réelle : mFDP et sa limite correspondante sont conservatrices, mais elles ne sous-estiment jamais la valeur du FDP. Enfin, mFDP est plus conservateur pour les petites valeurs de k, ce qui était également attendu sur la base des résultats de la simulation.

**CONCLUSION :**

Les gènes ayant la plus grande taille d'effet estimée peuvent avoir un PDR gonflé. Le problème survient lorsque des procédures de contrôle du FDR telles que BH sont combinées avec la procédure de double filtrage des Volcano Plots. Le contrôle du FDR sur un ensemble de découvertes n'implique pas le contrôle du FDR sur des sous-ensembles de ces découvertes.

L'inflation du FDR est élevée lorsque la variance des gènes différentiellement exprimée est inférieure à la variance des gènes non différentiellement exprimée. L'inflation du FDR peut se produire dans des contextes de données réelles. L'inflation du FDR est moins importante lorsqu’il y a de fortes corrélations ou une faible proportion de gènes nuls.

Le closed testing et le focused BH sont deux alternatives à la procédure BH et permettant un double filtrage tout en contrôlant le FDR.