TP "Multi-omics"

20 - 21 octobre 2020

Consignes

Vous avez jusqu'au 30 Octobre 2020 pour rendre le devoir. Vous devez nous remettre un fichier Rmd qui contient les réponses à toutes les questions. Vous inclurez également **toutes les commandes** qui vous ont permises de répondre aux questions.

N'oubliez pas d'inclure le nom et le prénom de tous les membres de l'équipe.

Vous pouvez nous joindre aux adresses suivantes:

- Arnaud Droit: Arnaud.Droit@crchudequebec.ulaval.ca
- Antoine Bodein: Antoine.Bodein@crchudequebec.ulaval.ca
- Charles Joly Beauparlant: Charles.Joly-Beauparlant@crchudequebec.ulaval.ca

Objectifs

Utiliser les méthodes vu en cours pour intégrer des données multi-omiques. Une grande partie du TP est réalisé grâce à la suite d'outils `mixOmics`. De l'aide est disponible sur leur site (http://mixomics.org/).

Partie I

0. Préparation

1. Chargez le package `mixOmics`
2. Téléchargez et importez les données (4 fichiers: `mirna.csv`, `mrna.csv`, `protein.csv`, `sample_group.csv`)

Question 1: Combien avez-vous d'échantillons ? de variables (mRNA, protéines, miRNA) ?

3. Le coefficient de variation est défini comme le rapport entre l'écart-type σ et la moyenne μ : $c_v = \frac{\sigma}{\mu}$ Construisez un fonction qui calcule le coefficient de variation à partir d'un vecteur.
4. A l'aide d'un histogramme `hist()` affichez la distribution de chacun des blocs.

Question 2: La distribution des coefficients de variation est-elle similaire dans les 3 blocs ? Si oui, quel type de donnée possède le plus de variabilité ?

5. Pour chacun des blocs, filtrez les données les plus variantes : $|c_v| \geq 0.15$

Question 3:: Combien reste-il de gènes ? de protéines ? de miRNA ?

Question 4: Quel est le gène le plus variant ? La protéine associé à ce gène est-elle présente dans le jeu de donnée.

Question 5: A l'aide des bases de donnée de votre choix répondez aux questions suivantes:

- Quel est le rôle de ce gène ?

- Sur quel chromosome est-il localisé ?
- Quelle est la longueur en nucléotide de sa séquence ?
- Quelle est la longueur en acides aminés de la protéine associée (ou des isoformes) ?

Partie II

1. Single-omic: l'ACP avec `mixOmics`

Question 6: A quoi sert l'Analyse en Composante Principale ? Expliquez brievement sont fonctionnement ?

1. Réaliser l'ACP sur les données mRNA.

Question 7: Combien de composantes retenez-vous ? Justifiez / Illustrez

2. Après avoir relancer l'ACP avec le bon nombre de composante, utiliser un graphique pour représenter les variables.

Question 8: Quelles sont les variables qui contribuent le plus à l'axe 1 ?

3. Avec un graphique, représenter les échantillons dans l'espace formé par les composantes. Les échantillons sont colorés en fonction du groupe. Affichez la légende et ajoutez un titre.
4. La *sparse ACP* `spca()` implémente une étape de *feature selection*. En utilisant la documentation de la fonction et/ou l'aide disponible en ligne, utilisez la `spca()` de manière a sélectionner 10 gènes sur la première composante et 5 gènes sur la seconde composante.

Question 9: Quelles sont les gènes que vous avez sélectionnés? *(une fonction est disponible)*

2. Projection on Latent Structures

1. Réalisez une PLS `pls()` avec les données mRNA et protéines en incluant 3 composantes `(ncomp = 3)`.

Question 10: A quoi sert la régression PLS pour l'intégration multi-omique?

2. Affichez un *scatter plot* des échantillons en affichant uniquement les composantes 2 et 3. Les échantillons doivent être coloriés par groupe. Ajoutez une légende et un titre.
3. Affichez un *arrow plot* en affichant uniquement les composantes 1 et 3. Les flèches doivent être coloriés par groupe. Ajoutez une légende et un titre.
4. La *sparse PLS* `spls()` implémente une étape de *feature selection*. En utilisant la documentation de la fonction et/ou l'aide disponible en ligne, utilisez la *sPLS* de manière a sélectionner (10 gènes, 9 protéines) sur la première composante, (5 gènes, 5 protéines) sur la seconde composante et (1 gène, 1 protéine) sur la troisième composante.

Question 11: Quels sont les variables sélectionnées sur la troisième composante.

5. Affichez un *CIM plot* à partir du modèle *sPLS*.

Question 12: Quels sont les gènes et les protéines les plus corrélés? Justifiez à partir de la matrice de corrélation calculée par `cim()`.

6. Toujours à partir du même modèle *sPLS*, affichez un *network plot* en affichant uniquement les les corrélations les plus forte ($\rho \pm 0.65$).

Question 13: Combien de clusters / sous-graphes observés vous ?

2. *multiblock* Projection on Latent Structures

1. Réalisez une multiblock PLS `pls()` avec les données mRNA, protéines et miRNA `(X = list(mrna, prot), Y = mirna)` en incluant 2 composantes `(ncomp = 2)`.

2. Comme la `spls()`, la `block.spls()` implémente une étape de *feature selection*. En utilisant la documentation de la fonction et/ou l'aide disponible en ligne, utilisez la fonction de manière a sélectionner (10 gènes, 9 protéines, 7 miRNA) sur la première composante et (5 gènes, 4 protéines, 3 miRNA) sur la seconde composante.

Question 14: Quels sont les variables sélectionnées sur la première composante.

3. Analyse supervisée : (s)PLS-DA

Le fichier `sample_groupe.csv` associe un groupe à chaque échantillon.

Question 15: Donnez la répartition des groupes.

1. Utilisez la `pls.da()` en utilisant les gènes (`X`) et le groupe (`Y`) avec 2 composantes.
2. Affichez le graphe des échantillons.

Question 16: Comparez ce graphe avec le graphe des échantillons obtenu avec l'ACP (1.3). Quel méthode permet d'obtenir de meilleurs clusters?

4. Analyse supervisée : block-(s)PLS-DA

1. Réalisez une multiblock sPLS-DA `block.splsda()` avec les données mRNA, protéines, miRNA `(X = list(mrna, prot, mirna))` et le groupe en incluant 5 composantes `(ncomp = 5)`.
2. Utiliser la fonction `perf()` sur le modèle obtenu.

Question 17: Quelle serait le nombre de composante minimal à inclure ?

3. Relancez le modèle avec 2 composantes et utilisez l'option `keepX` pour sélectionner 15 gènes, protéines et miRNA sur la première compoante et 10 gènes, protéines et miRNA sur la seconde composante.
4. Réalisez un *circos plot* avec le modèle obtenu en affichant les corrélations fortes $|\rho| > 0.5$. Ajoutez un titre.

Partie III

5. Mises en situation

Dans cette section, nous allons vous présenter deux designs expérimentaux et il vous faudra déterminer quelle est l'approche analytique à privilégier pour répondre aux questions demandées. Il s'agit d'intégrer à la fois l'informations sur l'analyse bioinformatique en partant des données brutes mais également de cibler les bonnes approches multiomiques.

1. Un de vos collègue s'intéresse aux effets de l'exposition à des polluants sur la santé des ours polaires. Pour ce faire, il a accès à des données transcriptomiques provenant d'une vingtaine de trios (un mère et sa portée de deux enfants) ainsi qu'à diverses mesures cliniques numériques pour tous les échantillons.
2. Vous travaillez sur un modèle murin et vous souhaitez comprendre les impacts d'un traitement sur le microbiote. Vous avez accès à des données de séquençage de type 16S ainsi qu'à des données de métabolomiques pour des souris traitées et pour des souris non-traitées. Vous pouvez prendre pour acquis que l'analyse primaire des données de métabolomiques a déjà été complétées et que vous avez déjà accès aux décomptes pour chaque molécules.