TP “Multi-omics”

2024-10-01

Consignes

Par groupe de 4 étudiants, on vous propose d'analyser des données multi- omiques issues de différents datasets. Vos objectifs seront d'identifiez une problématique liée aux données et vous analyserez celles-ci via les méthodes d'intégration vues en cours.

Vous présenterez vos résultats de vos analyses lors d'un exposé oral.

Dans les sections suivantes, vous trouverez les détails et points à aborder dans votre présentation.

Vous aurez besoin de R (>= 4.1) (https://cran.r-project.org; https://www.rstudio.com) Aini que les packages (`mixOmics`, `tidyverse`, `netOmics`, `igraph`, `gprofiler2`, `org.Hs.eg.db`, `org.Mm.eg.db`)

lien VM Rstudio: https://paris-rstudio.compbio.ulaval.ca/

Présentation des données:

Le dossier `Datasets/` contient les données suivantes ainsi que des publications associées:

- T2D: Diabète de type II, *Homo sapiens*, mRNA, protéines et cytokynes, variables cliniques = Insulino-Résistent/Sensible;
- Pregnancy: Suivi de grossesse normale, *Homo sapiens*, mRNA, protéines et cytokynes, variables cliniques = trimestre (t = 1;2;3) + post partum (t = 4);
- HeLa: Cellule HeLa, *Homo sapiens*, mRNA, protéines et translatomique, variables cliniques = 3 phases de la différentiation cellulaire.
- VitC: *Mus musculus*, mRNA, proteins, outcome = VitC level / sex / groups

Pour chaque dataset, l'article associé est situé dans le dossier `Papers/`.

Dans le dossier `Guides/` vous trouverez des tutoriels pour le preprocessing, l'utilisation de mixOmics, des graphes, des netOmics et des analyses d'enrichissement.

Plan d'analyse et slides à préparer

Présentation du dataset et de l'étude liée

- Contexte de l'étude
- Problématique
- Données dans l'étude et méthode d'analyse (se concentrer sur l'aspet intégratif).
- Brève description des résultats de l'étude

Importation des données en R et présentation du dataset

- Nombre d'échantillons
- Nombre de *features* (gènes, proteéines, ...)
- Nombre de classes (fichier info_sample)

Analyse préliminaire

Preprocessing

- Présentation des distributions des *features*, par échantillons, par classes, par bloc
- Au besoin, filterer les *features* peu abondants et avec peu de variabilité (objectif = avoir un dataset de petite taille pour la suite des analyses)
- Au besoin, log/scaler les données
- Discutez de vos choix

Mais aussi:

- Après vos filtres combien reste-il de *features* ?
- Quels sont les gènes les plus variants ? les protéines ? Quels sont leurs rôles biologiques ?
- Le gène le plus variant est-il traduit et présent dans le dataset ?

Analyse en Composante Principale

Pour chaque bloc:

- Faire une ACP (mixOmics) et discuter du nombre de composantes à inclure dans le modèle
- Présenter le graphe des variables, des échantillons
- Les données permettent-elles une bonne séparation des groupes ?
- Quelles variables contribuent le plus aux axes principaux ?

Analyse avec sPCA:

- Sélectionner 10 features sur la première composante et 5 sur la deuxième.
- Quelles sont les variables sélectionnées

Analyse supervisée

- Réaliser une PLSDA (Y = sample info)
- Comparer le graphe des échantillons de la PLSDA avec celui de la PCA.
- Lequel permet une meilleure séparation des classes ?

Analyse d'intégration

PLS

- Faire une première analyse PLS avec les gènes et les protéines
- Discuter du nombre de composantes à inclure dans le modèle
- Présenter le graphe des échantillons, des variables et arrow plot.

sPLS

- Faire une première analyse sPLS avec les gènes et les protéines (10 features/bloc composante 1, 5 pour composante 2).
- Quelles sont les variables sélectionnées ?
- Quels sont leurs rôles biologiques ?
- Y a t-il des fonctions biologiques enrichies ?

DIABLO

- Faire une première analyse combinée DIABLO en incluant les gènes, protéines et le groupe
- (Faire une seconde analyse en ajoutant le 3ième bloc)
- Optimiser le modèle (ncomp et keepX)
- Présentez circos plot et network plot de votre modèle final

Mise en réseau

Vous aurez besoin des données contenues dans le dossier `Supp/`

Réseaux de régulation de gène

- A partir des gènes filtrés, construisez un GRN avec netOmics
- Combien y a-t-il de noeuds, arête, noeuds déconnectés
- Affichez distribution des degrés
- Quel gène est le plus connecté ?

Réseaux PPI

- A l'aide de l'objet R contenant des informations PPI issues de BioGRID créer un réseau PPI avec seulement les protéines issues de votre dataset.
- Combien y a-t-il de noeuds, arête, noeuds déconnectés
- Affichez distribution des degrés
- Quelle protéine est la plus connectée ?

Connection gène-prot

Pour connecter le sous réseau gène avec le sous réseau protéine, vous devez utiliser à la fois l'information gène - protein coding pour l'information de traduction mais aussi l'information TF -> gène.

- Décrivez votre réseau.
- Combien de liens TF-gène et gène-protéine recensez-vous ?
- Faire une analyse de modularité, présentez les résultats
- A partir du noeud chosi avec soin, réalisez une analyse par random walk, justifiez le choix de la seed, présentez les résultats.
- Réalisez une analyse d'enrichissement sur les données

Conclusion

- Retour sur l'analyse:
  - Point forts et points faibles des outils
  - Difficultées rencontrées
  - Résumé des résultats