Master Bioinformatique, M2

Année 2025-2026

jsPCA : Identification rapide, évolutive et interprétable des domaines spatiaux et des gènes variables à partir de données transcriptomiques spatiales multi-coupes et multi-échantillons

Auteurs : Ines ASSALI, Paul ESCANDE et Paul VILLOUTREIX
Étudiant : Anaëlle JOO
Encadrant : Pr. Fiston-Lavier
UE : Projet bibliographique

La Transcriptomique Spatiale (TS) est une technologie récente qui permet de cartographier l'expression des gènes directement dans les tissus [2] : une coupe du tissu (*slice*) est réalisée, puis l'expression génique est mesurée sur chaque spot, dont la position spatiale est connue. La *joint spatial PCA* (jsPCA) est une méthode qui permet d'analyser les données de TS [1].

Prétraitement : La TS sort une matrice d'expression $X_j \in R^{n_j \times p}$ (n_j spots, p gènes) pour chaque coupe j. Cette matrice est ensuite normalisée durant une étape de prétraitement. Une matrice d'adjacence L_j est également créée à partir du graphe des k-voisins les plus proches des spots de la coupe.

Covariance spatiale : La matrice de covariance spatiale A_j est ensuite calculée à partir de la matrice d'expression génique X_j et la matrice d'adjacence L_j. Pour un jeu de coupes $\{(X_j, L_j)\}_j$, chaque matrice A_j est calculée, elle permet de comprendre au mieux la covariance de l'expression génique et l'auto-corrélation spatiale.

Eigen-décomposition : Les composantes principales spatiales sont les vecteurs propres communs à l'ensemble des matrices A_j. jsPCA permet l'analyse jointe de données de plusieurs coupes (*multi-slice*), puisqu'elle recherche un ensemble P_k de valeurs propres les plus grandes et communes aux matrices de covariances de chaque coupe. Les Gènes à Variabilité Spatiale (GVS) ont les plus grands coefficients absolus dans les premières composantes principales. Les données sont ensuite projetées sur les vecteurs propres en une représentation à dimension réduite $\widetilde{X}_j$.

Clustering : Un modèle de mélanges Gaussiens (GMM) est ensuite utilisé pour chercher les domaines spatiaux qui sont des régions montrant des motifs d'expressions similaires. Le nombre de domaines K peut être donné à priori (*"known ground truth"*) ou sélectionné via AIC/BIC : un post-traitement de raffinement spatial relabelise les spots dont la majorité des voisins a une étiquette différente (filtre de cohérence spatiale).

Résultats et performances : *jsPCA* a été évaluée sur deux jeux de données de référence : le cortex préfrontal humain (DLPFC) et l'atlas de développement embryonnaire de souris (MOSTA).

Sur le DLPFC, *jsPCA* atteint une précision de clustering supérieure (ARI = 0.62, NMI = 0.71), tout en utilisant bien moins de temps et de mémoire que des méthodes comme *SpatialPCA*, *BASS* ou *BayesSpace*.

L'analyse multi-coupes reste fiable jusqu'à 4 coupes, avec une baisse à 12 coupes due aux différences inter-échantillons, mais *jsPCA* reste compétitif en coût et en qualité. En validation *leave-one-out*, le NMI diminue légèrement mais l'ARI s'améliore, montrant que la méthode se généralise bien d'une coupe.

Sur MOSTA (jeu de données massif), *jsPCA* obtient les meilleures performances (NMI = 0.67, ARI = 0.48) et reste la seule méthode capable de traiter l'intégralité des coupes.

Enfin, *jsPCA* détecte efficacement les GVS, avec un chevauchement d'environ 80% avec SPARK-X.

Limites et perspectives Malgré ses performances, *jsPCA* présente certaines limites. Elle ne corrige pas explicitement les effets de lot (*batch effects*), ce qui peut affecter les analyses multi-échantillons. Cependant, sa modularité permet une intégration aisée avec des outils de correction existants en amont de l'analyse. Par ailleurs, bien que *jsPCA* s'étende naturellement à tout jeu de données spatialisées (spots ou cellules), son application à des données multi-omiques (protéomique, métabolomique) reste à développer.

Formulation mathématique

n	Nombre de spots
p	Nombre de gènes
$X_j \in R^{n_j \times p}$	Matrice d'expression de la coupe j
L_j	Matrice d'adjacence de la coupe j
$A_j = \frac{1}{2n_j} X_j^\top (L_j + L_j^\top) X_j$	Matrice de covariance spatiale de la coupe j
k	Nombre de composantes principales spatiales partagées
$P_k \in R^{p \times k}$	Matrice des vecteurs propres
D	Matrice diagonale des valeurs propres
$A_j \simeq P D_j P^\top$	Eigen-décomposition de la matrice de covariance spatiale
$\widetilde{X}_j = X_j P_k$	Projection des données sur les vecteurs propres
K	Nombre de domaines spatiaux

Glossaire

— *Gaussian Mixture Model* (GMM) : modèle probabiliste qui suppose que les données sont générées par un mélange de plusieurs distributions gaussiennes, chaque composante représentant un groupe ou domaine sous-jacent.

— *Adjusted Rand Index* (ARI) : indice qui mesure la similarité entre un clustering obtenu et une partition de référence. Il corrige l'accord attendu par hasard et varie de 0 (accord aléatoire) à 1 (accord parfait).

— *Normalized Mutual Information* (NMI) : mesure la quantité d'information partagée entre deux partitions, normalisée pour être comprise entre 0 (aucune information partagée) et 1 (correspondance parfaite).

— SPARK-X est un modèle pour l'identification des GVS. Il repose sur des tests statistiques adaptés aux structures spatiales et sert de référence pour évaluer la validité biologique des gènes détectés.

— *batch effects* : variation entre groupes d'échantillons en raison de facteurs non biologiques

Références

[1] Ines Assali, Paul Escande et Paul Villoutreix. « jsPCA : fast, scalable, and interpretable identification of spatial domains and variable genes across multi-slice and multi-sample spatial transcriptomics data ». In : *bioRxiv* (2025). DOI : 10.1101/2025.09.16.676466. eprint : https://www.biorxiv.org/content/early/2025/09/18/2025.09.16.676466.full.pdf. URL : https://www.biorxiv.org/content/early/2025/09/18/2025.09.16.676466.

[2] Vivien Marx. « Method of the Year : spatially resolved transcriptomics ». en. In : *Nat. Methods* 18.1 (jan. 2021), p. 9-14.