

Approche	Pour quoi faire ?		Points d'attention
WGCNA	<ul style="list-style-type: none"> <li>❖ Identifier des modules de gènes co-exprimés</li> <li>❖ Caractériser ces modules en lien avec des métadonnées cliniques (module corrélé à un trait phénotypique par exemple)</li> </ul>		<ul style="list-style-type: none"> <li>❖ Au moins 10 échantillons, optimalement <b>au moins 20</b> !</li> <li>❖ Si appliqué sur d'autres omiques que transcriptomique vérifier que : <ul style="list-style-type: none"> <li>➢ l'utilisation de la corrélation a du sens</li> <li>➢ l'assomption de soft-thresholding scale-free à du sens pour vos données</li> </ul> </li> <li>❖ Analyse signed (corrélations absolues) VS unsigned (corrélations avec signe) VS signed hybrid (corrélations positives)</li> </ul>
SNF	<ul style="list-style-type: none"> <li>❖ Identifier des groupes de patients présentant un profil multi-omique similaire</li> <li>❖ Identifier des sous-types phénotypiques</li> <li>❖ S'adapte a des données très multi-modales</li> </ul>		<ul style="list-style-type: none"> <li>❖ Tous les échantillons mesurés dans toutes les modalités</li> <li>❖ Maximum 20% de données manquantes par samples et par features (retirer sinon)</li> <li>❖ Utiliser des mesures de distance adaptées pour chaque modalité</li> </ul>
MOFA	<p>Extensions multi-omique de l'ACP.</p> <p>1. Identification des composantes captant un effet biologique d'intérêt.</p>	<ul style="list-style-type: none"> <li>❖ Uniquement non-supervisé</li> <li>❖ Intégration verticale de composantes communes (et éventuellement de spécifiques) à tous les omiques</li> <li>❖ Version adaptée au single-cell (MOFA+) permettant de faire de l'intégration horizontale et verticale</li> <li>❖ Gère tous types de données manquantes</li> <li>❖ Possibilité de faire de la sélection de variables</li> </ul>	<ul style="list-style-type: none"> <li>❖ Choix des distributions <ul style="list-style-type: none"> <li>➢ Si que des 0 et 1 dans le tableau ⇒ Bernoulli</li> <li>➢ Données de comptage ⇒ SizeFactor normalization + VST ⇒ Gaussian</li> <li>➢ Poisson moins utilisé (que si données: entiers positifs)</li> </ul> </li> <li>❖ Déséquilibre des données omiques en terme de nombre de variables (filtrer les plus gros blocs avec des filtres de variance).</li> <li>❖ MOFA+ est déjà intégré dans le package MOFA classique.</li> </ul>
RGCCA	<p>2. Enrichissement de ces composantes via leurs poids</p>	<ul style="list-style-type: none"> <li>❖ Supervisé ou non-supervisé</li> <li>❖ Intégration verticale uniquement</li> <li>❖ Possibilité de faire de la sélection de variables</li> </ul>	<ul style="list-style-type: none"> <li>❖ Faire une Cross-Validation (CV) lorsque RGCCA est utilisé en supervisée</li> <li>❖ Le déséquilibre en terme de nombre de variables est moins impactant, surtout lorsque RGCCA est utilisé en version supervisée.</li> <li>❖ Pour SGCCA, le nombre de variable sélectionnées par omique est plus flexible que MOFA mais doit être réglé par CV/permutations.</li> <li>❖ Pour les données de comptages, même normalisation que pour MOFA.</li> </ul>
Knowledge graphs	<ul style="list-style-type: none"> <li>❖ Modéliser des données multi-modales sous forme de graphe avec un vocabulaire de référence (ontologie)</li> <li>❖ Prendre en compte ≠ niveaux de généralisation</li> <li>❖ Compléter des données graphes avec des règles logiques</li> <li>❖ Enrichir des données expérimentales avec des connaissances publiques</li> </ul>		<ul style="list-style-type: none"> <li>❖ SPARQL produit des matrices en résultat de requête mais ne fait pas d'analyse stat → cf MOFA, RGCCA, SNF, WGCNA</li> <li>❖ SPARQL peut produire des sous-graphes mais ne fait pas d'analyse de la topologie ni des parcours → cf networkX, Cytoscape ...</li> </ul>