Evaluation des méthodes de réduction de dimension pour la détection de cancers de phénotype intermédiaire à partir de données génomiques

Poizat Jérôme

Supervisé par :

Dr. Nicolas Alcala, Dr. Lynnette Fernandez-Cuesta et Dr. Matthieu Foll

International Agency for Research on Cancer

World Health Organization

bioinfo@lyon

BIOINFORMATIQUE MOLÉCULAIRE: MÉTHODES & ANALYSES

UNIVERSITÉ DE LYON

Université Claude Bernard Lyon 1

2019

Résumé

Dans le contexte de la classification moléculaire des cancers, pouvoir identifier des cancers de phénotype intermédiaire est un enjeu important pour la médecine personnalisée auquel on peut répondre grâce la croissance exponentielle des données génomiques. Le principal défi dans l'analyse des données génomiques est l'intégration et l'interprétation des données de manière à apporter des renseignements pertinents d'un point de vue biologique.

L'ACP et UMAP sont deux méthodes de réduction de dimension qui permettent d'analyser des données génomiques. Pour explorer la capacité de l'ACP et d'UMAP à identifier efficacement des cancers rares de phénotype intermédiaire, et explorer les limites de ces deux méthodes, des données de séquençage d'ARN ont été simulées à partir d'un modèle mettant en scène des échantillons de deux cancers et des échantillons de phénotype intermédiaire. Simuler des données permet de connaitre leur réalité et donc d'évaluer la capacité des méthodes à modéliser cette réalité et correctement représenter le continuum entre deux cancers. Le R^2 est une métrique utilisée pour quantifier l'écart entre les données observées et théoriques, et permet ainsi d'évaluer la justesse de la position des échantillons intermédiaires dans la représentation graphique à dimension réduite. Avec cette méthode d'évaluation, il est possible de faire l'exploration des caractéristiques des données pouvant influencer l'identification ou non des cancers intermédiaires.

Nos résultats ont montré que, dans le contexte où l'on cherche à identifier des cancers de phénotype intermédiaire entre deux groupes de cancers, l'ACP est un algorithme de réduction de dimension efficace à cette tâche. UMAP en revanche est un algorithme qui dans ce contexte a une efficacité moindre dans la représentation des cancers de phénotype intermédiaires, notamment lorsque le nombre d'échantillon est faible.

Table des matières

Table des figures

Abréviations

IARC : International Agency for Research On Cancer

MPM : Malignant Pleural Mesothelioma

ACP : Analyse Composante Principale

CP : Composante Principale

UMAP : Uniform Manifold Aproximation Projection

t-SNE : t-distributed Stochastic neighbor embedding

Logiciels

R 3.5.2
Packages R : ade4 1.7, umap 0.2.0.0

Introduction

Présentation de l'IARC et du groupe GCS

Fondée en 1965, le Centre International de Recherche sur le Cancer (IARC *: International Agency for Research On Cancer*) est une agence intergouvernementale spécialisée en oncologie faisant partie de l'Organisme Mondiale de la Santé (WHO : *World Health Organization*). Depuis sa fondation, l'IARC est situé à Lyon, France. L'objectif principal de l'IARC est d'encourager la collaboration internationale pour la recherche sur le cancer. Dans ce contexte international, l'agence mène de nombreuses études sur les causes des cancers et les mesures préventives à entreprendre, en rassemblant des experts de nombreuses disciplines scientifiques : épidémiologie, biologie humaine, biostatistiques et bioinformatique.

Vue de la tour de l'IARC (2018)

Le *Genetic Cancer Susceptibility Gro*up (GCS) est un groupe de recherche de l'IARC qui se focalise sur l'identification de variants génétiques contribuant à la susceptibilité au cancer, mais poursuit également des recherches sur les altérations génomiques qui se produisent durant la carcinogénèse et pendant la progression tumorale. Pour mener ces recherches, l'équipe a recours à des techniques avancées de génomique et de bio-informatique.

Contexte de l'étude : la classification des cancers

La classification des cancers est d'une importance capitale en oncologie, celle-ci a pour but de prévoir le prognostique d'une tumeur, et de permettre l'utilisation ou le développement de thérapeutiques adaptées. Jusqu'à récemment, une majorité des classifications des cancers a été établie par les médecins pathologistes à partir des caractéristiques morphologiques des cellules cancéreuses.

Cependant, la « révolution » génomique née de la maturation des techniques de séquençage à haut débit est en train de bouleverser les pratiques. Le développement de ces technologies de séquençage a contribué à de grands progrès dans la recherche en biologie et en médecine. En cancérologie le séquençage de tumeurs devient une pratique repandue qui permet le développement de nouvelles méthodes plus efficaces de diagnostic, de meilleurs suivis et prise en charge des maladies, et de nouveaux traitements mieux ciblés. Le premier enjeu de ces nouvelles technologies de séquençage en cancérologie est de rendre le génotypage des tumeurs facile et accessible, et de fortifier les bases de données pour alimenter la recherche sur le cancer. Les classifications moléculaires de tumeurs sont principalement faites à partir de données de séquençage d'ARN ou de données de méthylation d'ADN, car ces deux sources de données sont celles qui permettent de mieux caractériser les tumeurs (Katherine A. Hoedley *et al.,* 2018).

Les données génomiques ont le potentiel d'informer une classification de haute définition qui permettrait une prise en charge clinique de précision. Il existe actuellement des initiatives de reclassification des cancers à partir de données génomiques avec des techniques d'apprentissage non supervisée. L'exemple le plus probant est le projet de reclassification des 10 000 tumeurs de l'initiative TCGA (The Cancer Genome Atlas) qui a permis de mettre en avant les liens existants entre les différents types de cancers et de découvrir de nouveaux groupes de tumeurs avec des cibles thérapeutiques uniques (Katherine A. Hoedley *et al.,* 2018).

L'initiative « rare cancer genomics » du groupe GCS a pour but de fournir une telle classification pour les tumeurs rares, lesquelles sont les oubliées des grandes initiatives comme TCGA, un tel projet récent de classification a été effectué pour le Malignant Pleural Mesothelioma (MPM, mésothéliome pleural malin en français). A partir d'analyses non supervisées de données génomiques, l'étude suggère que le profile moléculaire et le pronostique de cette maladie sont mieux expliqué par un continuum de classes moléculaires de tumeur plutôt que par la classification discrète actuelle de la WHO en trois classes, qui est la référence actuelle pour la prise en charge de la maladie dans le monde (Nicolas Alcala *et al.,* 2018). Cette étude est un exemple qui met en avant le fait que les classifications discrètes actuelles peuvent simplifier la réalité complexe des cancers, et qu'il existe en pratique des cancers de phénotype intermédiaire, qui sont difficiles à ranger dans des catégories discrètes. Une classification pas assez précise peut rendre le diagnostic d'une condition atypique difficile et donc avoir un impact sur la prise en charge du patient. Affiner les classifications actuelles avec des technologies de machine learning est alors nécessaire pour améliorer la prise en charge des patients en s'orientant sur une stratégie de médecine personnalisée.

La classification moléculaire des tumeurs : un exercice de classification non-supervisée

Un enjeu majeur des banques de données génomiques est, à partir d'un volume de données très large, de faire des analyses et des interprétations pertinentes à l'aide des technologies de *Data Sciences*, pour d'abord réduire l'espace dimensionnel des données, afin de les rendre visualisables et interprétables, et de par la suite faire des analyses non supervisées pour extraire de l'information dans les données, par exemple découvrir de nouveaux modèles, des groupes (clustering), des patterns, ou des corrélations.

Clustering

Le clustering est une technique statistique d'analyse de données consistant à former des groupes de données de manière à ce que les éléments dans un groupe (cluster) soient plus similaires entres eux et moins similaires aux éléments d'autres groupes. Autrement dit le but du clustering est de classer les données en regroupant ceux présentant des propriétés similaires. Cette technique de classification non supervisée est populaire et utilisée dans divers domaines dont le machine learning, l'analyse d'images, la finance et le marketing, la compression de données et bio-informatique. Le processus de clustering est souvent précédé d'une étape de réduction de la dimension des données pour que la classification des données soit faite en prenant en compte un maximum d'information signal et un minimum de bruit. Il existe une grande diversité d'algorithmes de clustering qui utilisent des approches différentes pour classifier les données. Le choix de l'algorithme de clustering se fait souvent expérimentalement et dépend des caractéristiques des données et du type de cluster recherché (Pranav Nerurkar *et al.,* 2018).

Réduction de dimension

Les données génomiques sont dans des espaces de grande dimension. Le nombre de dimensions est généralement de l'ordre de grandeur du nombre de gènes observés. Ce très grand espace dimensionnel rend presque impossible l'interprétation et la visualisation des données à l'état brut, entre autres à cause de la difficulté à représenter des espaces de grande dimension et du phénomène de la « malédiction de la dimension » qui est le fait que les distances tendent à être identiques dans un grand espace dimensionnel. Il est alors nécessaire de réduire la dimensionalité, c'est à dire réduire le nombre de variables dans les données en prenant soin d'éliminer les variables de « bruit », qui ne contiennent pas d'information biologique pertinentes, tout en conservant le maximum de « signal » biologique, afin de rendre possible la visualisation, l'interprétation et l'exploration des données. L'Analyse en composante principale (ACP) est la technique de réduction de dimension la plus populaire.

Analyse de composante principale (ACP)

L'ACP est une des techniques les plus utilisés de réduction de dimension pour l'analyse de données multivariées. Cette méthode est employée dans divers domaines d'applications tel que l'astronomie, la physique, la finance quantitative, neurosciences, la biologie et bio-informatique. Cette technique statistique d'apprentissage non supervisé réduit la dimensionnalité des données tout en conservant le plus de variabilité possible, et permet d'observer la structure des données. Ceci est fait via des projections géométriques des données sur un nombre de dimensions plus faible qu'à l'origine. L'ACP engendre des composantes principales (CP), qui sont des variables linéairement non corrélées qui retiennent le maximum de variabilité dans les données. En retenant le maximum de variabilité on conserve l'essentiel de l'information dans les données. La première composante principale explique la plus grande variance dans les données, et les composantes suivantes expliquent le maximum de la variance restante (Lan T. Jolliffe *et al.,* 2016). L'ACP permet ainsi de visuellement identifier des instances aberrantes dans les données ou d'identifier des groupes. Un exemple en génomique peut être l'identification de groupes d'échantillons correspondant aux différentes origines tissulaires.

Le principal avantage de l'ACP par rapport à d'autres méthodes de réduction de dimension non linéaire est que les axes, qui sont les CP, ont une échelle interprétable correspondant au taux de variance expliquée, autrement dit à la concentration d'information. Un inconvénient de l'ACP est la perte d'information dans la sélection des CP, puisque celles-ci sont le résultat de projection des données. D'autre part l'ACP ne peut que capturer des structures linéaires dans les données, et analyse les données seulement dans leur structure globale (Jake Lever *et al.,* 2017) (***Figure 1***).

Uniform Manifold Approximation and Projection (UMAP)

UMAP est une technique récente de réduction de dimension non linéaire et de pré-clustering fondée sur les concepts de la géométrie du manifold et de l'algèbre topologique. Comme l'ACP, UMAP permet de visualiser des données de haute dimension dans un espace de dimension réduite, cependant UMAP a l'avantage de ne pas faire d'hypothèse de linéarité. Le travail de Alex Diaz-Papkovich *et al.* (2018) sur les données du projet UK Biobank est un exemple d'utilisation d'UMAP qui a permis d'observer visuellement le continuum génétique entre différentes populations ethniques (***Figure 2***). Un autre exemple d'utilisation de UMAP est la visualisation de données single-cell par Etienne Becht *et al.* (2019). Cet algorithme est souvent adressé comme étant un potentiel successeur de l'algorithme t-SNE (t-distributed Stochastic neighbor embedding) qui est l'algorithme le plus couramment utilisé pour les taches de réduction de dimension non linéaires. UMAP et t-SNE sont deux

algorithmes de réduction de dimension qui ont la particularité de mettre en avant les structures locales dans les données, en opposition aux méthodes linéaires, comme l'ACP, qui capturent seulement la structure globale (***Figure 1***). UMAP à l'avantage par rapport à t-SNE de conserver plus d'information sur la structure globale, tout en étant beaucoup plus rapide (Leland McInnes *et al.,* 2018).

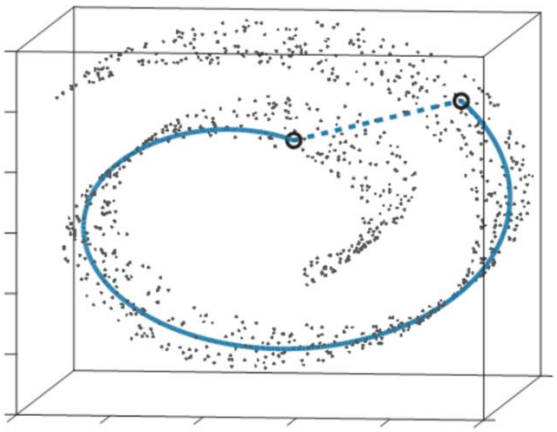

Figure 1 : Points dans un espace à 3 dimensions, représentation des distances locales et globales
Les points encerclés sont proches dans l'espace euclidien (distance en pointillé) mais éloignées dans l'espace des données (distance en trait pleins). Si l'on cherche à réduire la dimension de cet espace, la PCA, qui capture la structure globale des données, ne permet pas de visualiser la distance en trait pleins mais seulement la distance en pointillé. Au contraire, UMAP qui capture la structure locale des données permet de visualiser la distance en trait pleins mais perd en partie l'information des distances globales (distance en pointillé).

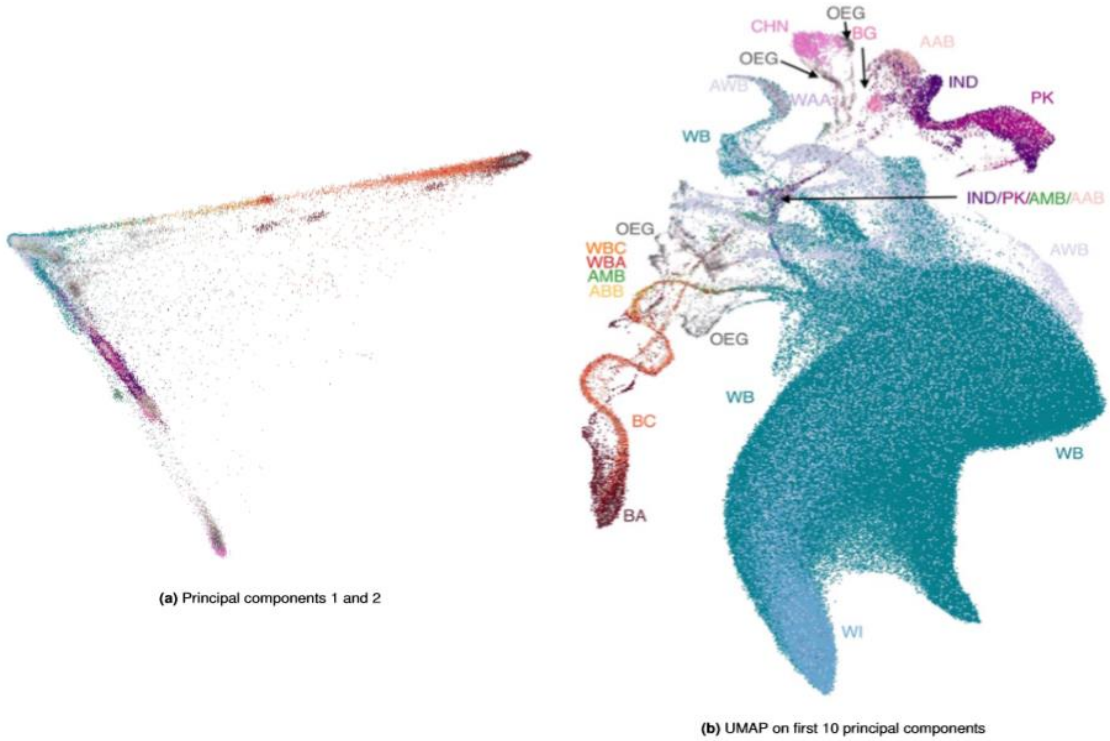

Figure 2 : Représentation graphique des données *UK Biobank* (Alex Diaz-Papkovich *et al., 2018*)
Représentation graphique des données *UK Biobank* avec L'ACP (a) et UMAP (b). UMAP permet une meilleure visualisation des structures locales des données tout en conservant la structure globale et la totalité de l'information.

Objectif du stage : caractériser les méthodes de réduction de dimension pour la détection de cancers de phénotype intermédiaire

Dans le contexte de la classification moléculaire des cancers, pouvoir identifier des cancers de phénotype intermédiaire est un enjeu important pour la médecine personnalisée auquel on peut répondre grâce la croissance exponentielle des données génomiques. Le principal défi dans l'analyse des données génomiques est l'intégration et l'interprétation des données de manière à apporter des renseignements pertinents d'un point de vue biologique.
La question est alors de savoir comment identifier efficacement les cancers rares de phénotype intermédiaire, c'est-à-dire de discerner quelles méthodes de machine learning non supervisée permettent d'obtenir une bonne représentation de la continuité entre les cancers, et identifier les limites de ces méthodes. Il est également important de savoir dans quelles conditions les données sont propices à l'identification de cancer intermédiaires.

Pour répondre à cette question, j'ai exploré les méthodes de réduction de dimension les plus classiquement utilisées : l'ACP et UMAP. L'ACP est une méthode de décomposition linéaire historiquement utilisée pour la réduction de dimension et UMAP est un algorithme non-linéaire moderne. Dans le cas de UMAP, l'exploration de la méthode consiste aussi à l'exploration des paramètres de l'algorithme, c'est-à-dire découvrir quelles valeurs de ces paramètres permettent une identification efficace des cancers intermédiaires. En parallèle de la caractérisation des méthodes de réduction de dimension, est fait l'exploration des caractéristiques des données pouvant influencer l'identification ou non des cancers intermédiaires. Ces caractéristiques sont par exemple la taille des données, la proportion de cancer intermédiaire dans les données, et la variabilité entre les cancers.

Pour caractériser ces méthodes de réduction de dimension, j'ai effectué des simulations de données RNAseq, de manière à avoir deux groupes d'échantillons de cancer et des intermédiaires. Le but étant de connaitre la réalité des données simulé pour évaluer la capacité des méthodes à modéliser cette réalité et correctement représenter des intermédiaires entre deux cancers.

Matériels et méthodes

Approche par simulation de données RNAseq

Pour l'exploration de la capacité de l'ACP et d'UMAP à identifier efficacement les cancers rares de phénotype intermédiaire, nous avons choisis de simuler des données de séquençage d'ARN. Simuler des données permet de connaitre leur réalité afin de pouvoir évaluer la capacité des méthodes à modéliser cette réalité et correctement représenter le continuum entre deux cancers. A contrario, dans les données réelles, on ne sait jamais avec certitude si un échantillon constitue un phénotype intermédiaire, et il est donc difficile d'évaluer la capacité d'une méthode à les détecter.

Modèle

J'ai participé à la conception d'un modèle in silico de tumeurs permettant de représenter deux groupes moléculaires I et II distincts de tumeurs et un nombre spécifié de phénotypes intermédiaires dont le profil moléculaire est intermédiaire entre I et II. Dans ce modèle, nous supposons que *n* tumeurs venant d'individus distincts ont été séquencées. Parmi ces *n* tumeurs, une proportion α sont des phénotypes intermédiaires (soit α x *n* échantillons), et nous supposons que les deux groupes moléculaires ont une même proportion d'échantillons (1 - α) /2 (soit (1 - α)*n/2 échantillons). Nous supposons que l'expression a été quantifiée pour un nombre *m* de gènes pour chaque échantillon. Parmi ces gènes, nous supposons qu'une quantité m_1 est différentiellement exprimée, c'est-à-dire que des gènes correspondent à des voies métaboliques qui présentent des différences entre les deux groupes de tumeurs. Nous dénotons $\mu_1 \, et \, \mu_2$ l'expression moyenne de ces gènes dans les deux groupes, respectivement, et σ leur variance. Une proportion $m\text{-}m_1$ de gènes est en revanche similairement exprimée entre les deux groupes, représentant tous le gène correspondant à des voies métaboliques qui sont semblables entre les deux groupes de tumeur (par exemple assurant des fonctions essentielles de la cellule non liée au cancer). Les échantillons de phénotype intermédiaire ont une moyenne d'expression μ_3 de leurs gènes différentiellement exprimés de sorte qu'elle soit comprise entre les deux moyennes d'expression des gènes des groupes de cancers $\mu_1 \, et \, \mu_2$. Chaque échantillon intermédiaire possède un coefficient de mixture β qui est un nombre compris entre 0 et 1 qui indique à quel point l'échantillon est proche du premier cancer (β = 0) ou du deuxième cancer (β = 1). Par exemple un échantillon intermédiaire équidistant aux deux cancers possède un shift factor de 0,5. μ3 est alors calculée de la façon suivante :

$$\text{Si } \mu_1 > \mu_2 \text{ alors } \mu_3 = \mu_1 - |\mu_1 - \mu_2| * \beta$$

$$\text{Sinon } \mu_3 = \mu_1 + |\mu_1 - \mu_2| * \beta$$

Simulation

Les données du modèle sont représentées par une matrice d'expression, avec les gènes fictifs en colonne, et les échantillons fictifs en ligne (***Figure 3***). Les données RNAseq simulées ont été généré avec R 3.5.2, à l'aide d'un script que j'ai réalisé.

Le nombre total de gènes *m* est fixé à 6 000, et le nombre de gènes avec des moyennes d'expressions différentes m_1 est fixé à 2 000, ces nombres sont du même ordre de grandeur que ce qui est observée expérimentalement dans les données de séquençage d'ARN après sélection des gènes expliquant le plus de variance (Nicolas Alcala *et al.*, 2018). Pour chaque gène, l'expression dans les échantillons suit une distribution normale avec une moyenne choisie aléatoirement et une variance σ définie de sorte que le rapport $\frac{|\mu_1 - \mu_2|}{\sigma}$ soit constant. Ce rapport reflète à quel point les cancers sont distants d'un point de vue moléculaire. Pour chaque échantillon, le coefficient de mixture β est choisi aléatoirement selon une distribution uniforme entre 0 et 1.

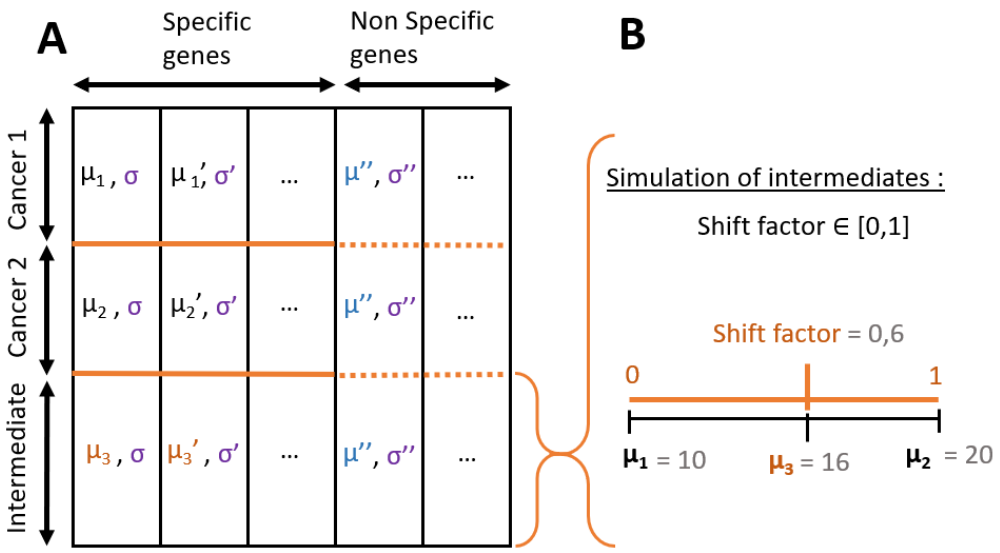

Figure 3 : Schéma récapitulatif de la simulation des données RNAseq

A : Matrice d'expression génétique avec les moyennes d'expressions (μ) et les variances (σ).

B : Exemple de calcul de la moyenne d'expression d'un intermédiaire avec un coefficient de mixture de 0,6.

Réduction de dimension

Après avoir généré la matrice d'expression génétique, les données simulées sont traitées par les algorithmes de réduction de dimension ACP et UMAP avec les packages R correspondant (ade4 1.7 et umap 0.2.0.0). Ces algorithmes retournent les coordonnées spatiales des échantillons dans un espace de dimensions. Pour l'ACP les deux dimensions correspondent aux deux premières CP.

Evaluation des méthodes de réduction de dimension

Puisque les données sont simulées, la réalité des données est connue. Autrement dit, pour chaque échantillon il est possible de dire avec certitude s'il appartient au premier cancer, au deuxième cancer, ou s'il s'agit d'un intermédiaire. L'objectif est alors de mesurer à quel point les réductions de dimensions produites par ces deux algorithmes rendent compte de la réalité des données. Plus particulièrement nous voulons savoir si les échantillons intermédiaires sont bien placés à des positions « intermédiaires », c'est-à-dire entre les deux clusters des cancers.

Pour évaluer à quel point un intermédiaire est correctement représenté dans la réduction de dimension, la position spatiale de l'échantillon entre les deux clusters est comparée à son coefficient de mixture (β). Il est attendu dans une bonne représentation qu'un échantillon avec un β proche de 0 soit graphiquement représenté proche du cluster du premier cancer, et qu'un échantillon avec un β proche de 0,5 soit graphiquement représenté à mi-chemin entre les deux clusters de cancer. Une façon simple de mesurer la position d'un échantillon relatif aux deux clusters est de projeter ce point sur la droite passant par les centroïdes des deux groupes, et de normaliser la distance entre les deux centroïdes à 1. Il est alors attendu dans une bonne représentation qu'un échantillon avec un β proche de 0,5 soit projeté à une distance de 0,5 sur le segment reliant les deux centroïdes (***Figure 4***). Il suffit ensuite de comparer les distances des projections et les β associés pour évaluer si la réduction de dimension positionne correctement les intermédiaires.

Le coefficient de détermination (R^2) est une métrique qui permet de quantifier à quel point les données observées, ici la distance des projections, sont ajusté aux données théoriques, ici le coefficient de mixture. Le R^2 peut prendre des valeurs comprises entre 0 et 1, un R^2 faible indique un écart important entre les données observées et théoriques, à l'inverse un R^2 proche de 1 signifie que les données observées sont ajustées aux données théoriques. Calculer le R^2 des distances des projections contre les coefficients de mixture associés permet d'avoir une métrique qui évalue la capacité des méthodes de réduction de dimension à correctement représenter les intermédiaires.

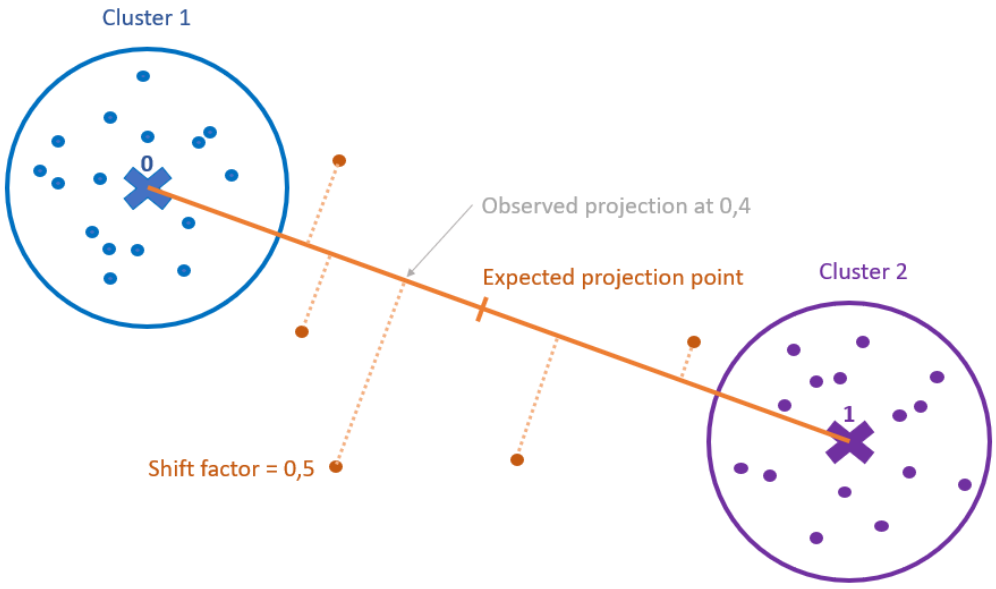

Figure 4 : Schéma récapitulatif de l'évaluation de la position des intermédiaires avec les projections
Pour mesurer la position d'un échantillon relatif aux deux clusters, le point est projeté sur la droite passant par les centroïdes. La distance entre les deux centroïdes est normalisé à 1. Il est alors attendu dans une bonne représentation qu'un échantillon avec un coefficient de mixture proche de 0,5 soit projeté à une distance sur la droite de 0,5.

Exploration de l'espace des paramètres

Le R^2 des projections est une métrique qui permet d'évaluer la justesse de la position des échantillons intermédiaires dans la représentation graphique à dimension réduite. Avec cette méthode d'évaluation, il est possible de faire varier des paramètres et savoir si ceux-ci ont un effet sur la justesse de la représentation graphique. De la même manière on peut déterminer pour quelles valeurs de paramètres la représentation des intermédiaires est optimale.

Les paramètres intéressants à explorer, c'est-à-dire ceux pouvant avoir un impact sur l'identification des intermédiaires, sont divisés en deux catégories. D'une part il y a les paramètres dépendants des données, qui sont la taille des données (*n*), la proportion d'intermédiaires (α), et la distance entre les deux clusters déterminée par le rapport $\frac{|\mu_1 - \mu_2|}{\sigma}$ où μ_1 et μ_2 sont les moyennes d'expression d'un gène dans les deux clusters et σ la variance de l'expression.
D'autres part il y a les paramètres liés aux techniques de réduction de dimension. Pour l'ACP il n'y a pas de paramètres à explorer. Pour UMAP il existe de nombreux réglages possibles, les plus intéressants à explorer sont les paramètres *n_neighbours* et *min_dist*, les autres paramètres d'UMAP ayant un impact marginal sur la réduction de dimension (Leland McInnes *et al., 2018*). *n_neighbours* est le nombre de plus proche voisins utilisé par UMAP pour l'apprentissage de la structure locale des données. *min_dist* est la distance minimal autorisé entre deux points dans la représentation graphique de UMAP, des valeurs plus élevé de *min_dist* empêche UMAP de trop grouper les points et permet de mieux conserver la structure topologique des données.
Pour chaque combinaison de valeurs des paramètres, 100 simulations sont réalisées dans le but de d'avoir la distribution des R^2.

Résultats

Exploration de l'espace des paramètres avec l'ACP

Pour toutes les combinaisons de valeurs de *n*, α, et $\frac{|\mu_1 - \mu_2|}{\sigma}$, les valeurs de R^2 obtenues avec l'ACP sont globalement compris entre 0,8 et 1, ce qui indique que les données observées sont ajustées aux données théoriques. Le seul cas où le R^2 est inférieure à 0,8 est quand *n*, α *et* $\frac{|\mu_1 - \mu_2|}{\sigma}$ sont petits, autrement dit quand le nombre d'échantillon est faible ($n \leq 100$) la proportion d'intermédiaires est faible (α = 2%) et les cancers sont moléculairement proches ($\frac{|\mu_1 - \mu_2|}{\sigma}$ = 0,5) (***Figure 5***).

Exploration de l'espace des paramètres avec UMAP

Dans notre modèle, les valeurs de R^2 obtenues avec UMAP sont globalement plus faibles (entre 0,3 et 0,8) que celles obtenues avec l'ACP. Les R^2 sont les plus faibles (≈ 0,3) lorsque *n* et α sont petits. Une augmentation des valeurs de *n* et α a pour conséquence une augmentation du R^2. Ce qui signifie que généralement les données sont mieux représentées lorsque le nombre d'échantillons et la proportion d'intermédiaire sont grands. Les valeurs de $\frac{|\mu_1 - \mu_2|}{\sigma}$ qui donnent des $R^2 > 0,9$ sont $\frac{|\mu_1 - \mu_2|}{\sigma}$ = 1 et $\frac{|\mu_1 - \mu_2|}{\sigma}$ = 2 ce qui signifie dans le modèle que des meilleurs R^2 sont obtenus lorsque les cancers ne sont ni moléculairement trop proches, ni moléculairement trop éloignés. En ce qui concerne le paramètre *min_dist* de UMAP, de plus fortes valeurs de *min_dist* (*min_dist* = 2.5) sont associés à des valeurs de R^2 plus proche de 1 (***Figure 5***). Graphiquement cela se traduit par une meilleure visualisation

de la continuité des intermédiaires entre les deux clusters de cancer, une plus grande valeur du paramètre *min_dist* permet une meilleure visualisation cette continuité (***Figure 6***). Ceci est valable pour des valeurs de *min_dist* inférieures à 2,5.

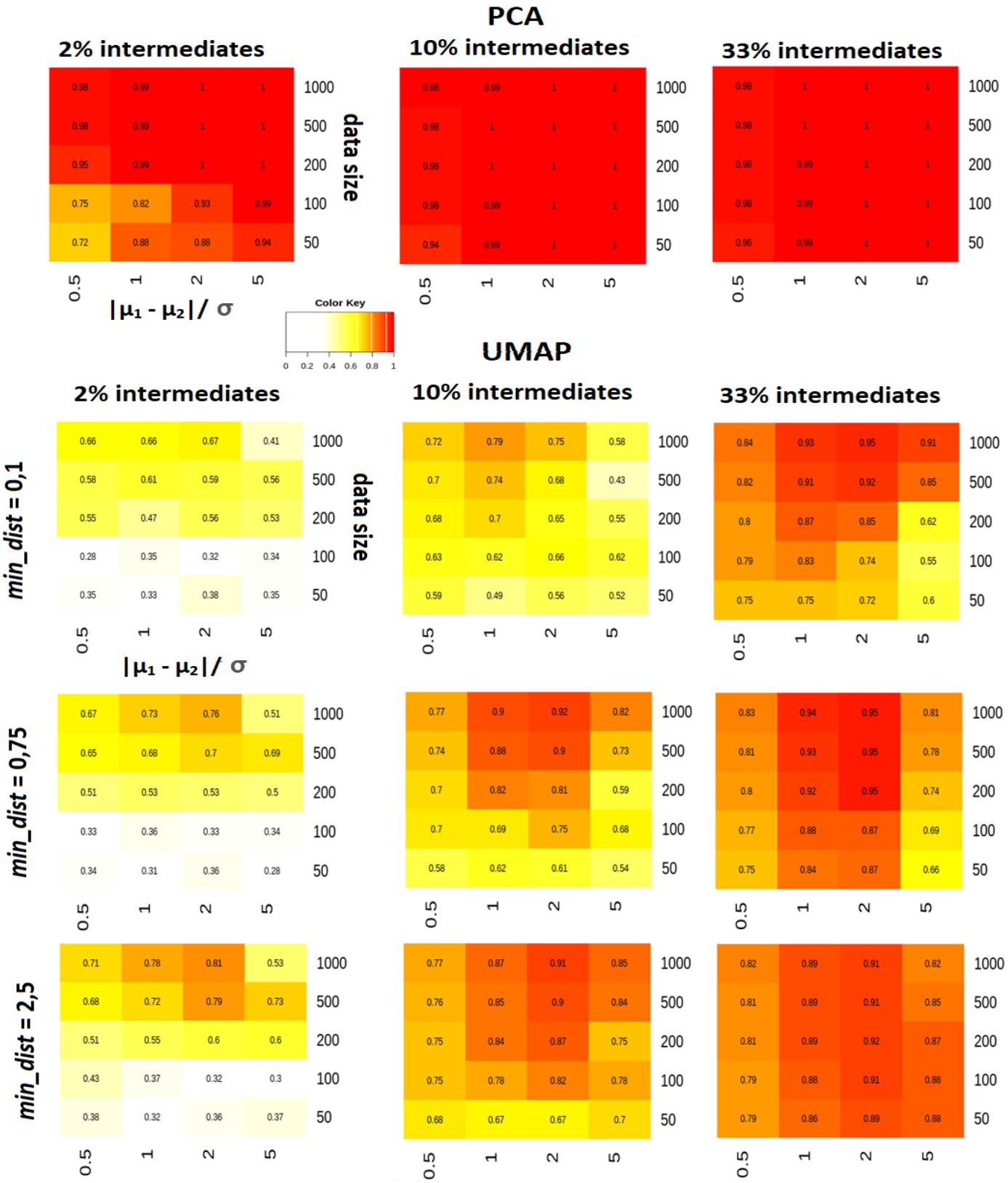

Figure 5 : Moyenne des R^2 obtenus avec l'ACP et UMAP dans l'exploration de l'espace des paramètres

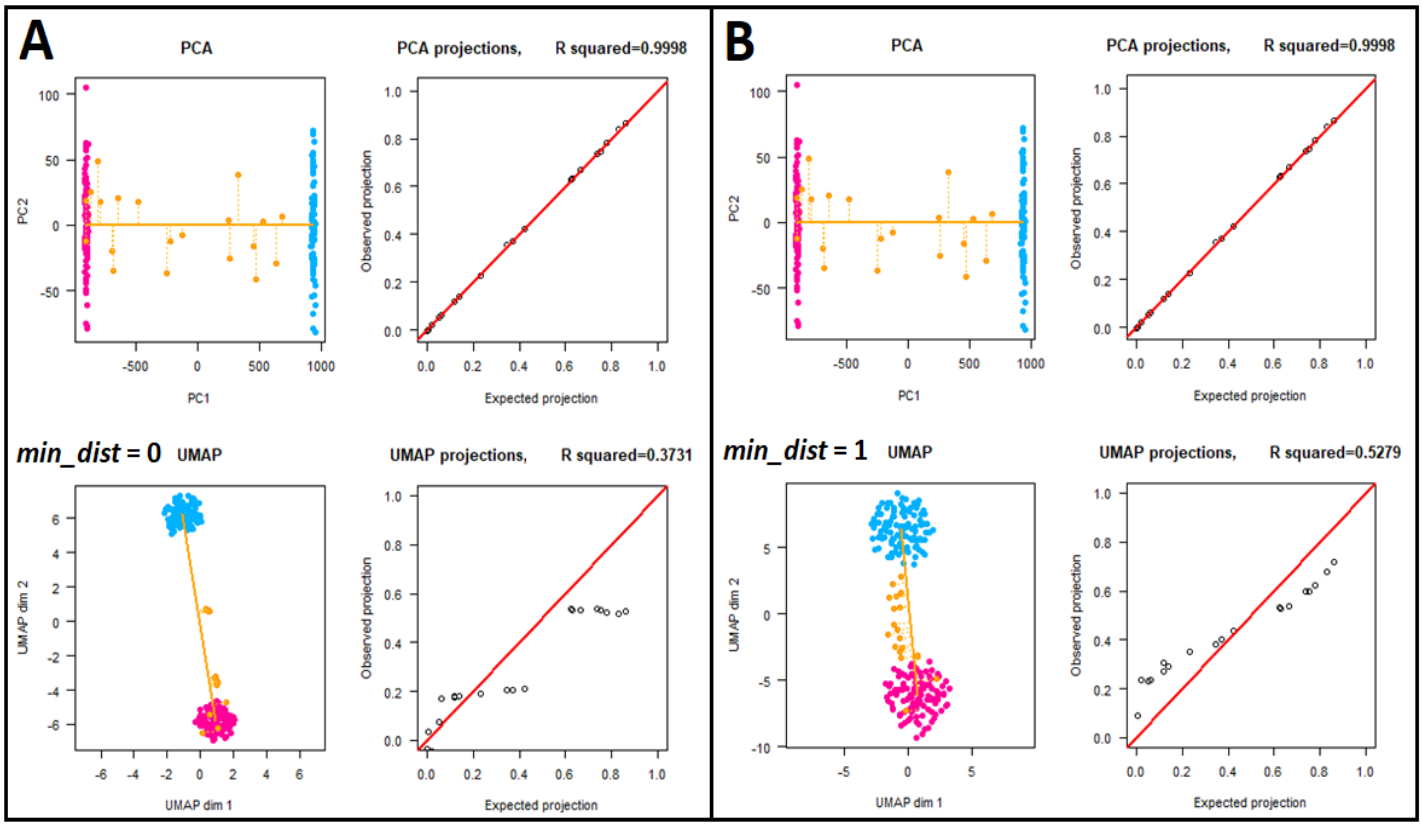

Figure 6 : Effet du paramètre *min_dist* de UMAP, réduction de dimension de données simulées avec l'ACP et UMAP, avec le graphique des projections observées en fonction des projections théoriques

En bleu et violet sont représentées les échantillons des deux groupes de cancers, et en orange les intermédiaires. **A** : Avec une valeur de *min_dist* = 0, UMAP donne une fausse représentation du continuum d'intermédiaires dans les données. **B** : Avec une valeur de *min_dist* = 1, UMAP donne une meilleure représentation du continuum d'intermédiaires dans les données.

Comportements observés avec UMAP

Lors des réductions de dimension des simulations du modèle avec UMAP, certains comportements récurrents ont été observées. Ces comportements ont un effet non négligeable sur la justesse de la représentation des données.

Une tendance d'UMAP est de grouper les points, un « effet de gravité » qui « attire » les points vers les régions les plus dense, créant un biais dans la représentation. Dans le cas de notre modèle cela a pour effet de soit grouper les intermédiaires dans les clusters des groupes des cancers (***Figure 7A***), soit de représenter un troisième cluster intermédiaire (***Figure 7B***). De ce fait la représentation de UMAP est sensible aux régions de plus faible densité qui dans le cas de notre modèle, fait des « trous » dans le continuum d'intermédiaires (***Figure 7C et D***). Ces effets sont moins marqués lorsque l'on augmente le paramètre *min_dist* (***Figure 6***).

Une autre tendance marginale de UMAP est de représenter le continuum d'intermédiaires sur un arc de cercle et non une ligne droite lorsque le nombre d'échantillons intermédiaire est élevée (***Figure 7D***).

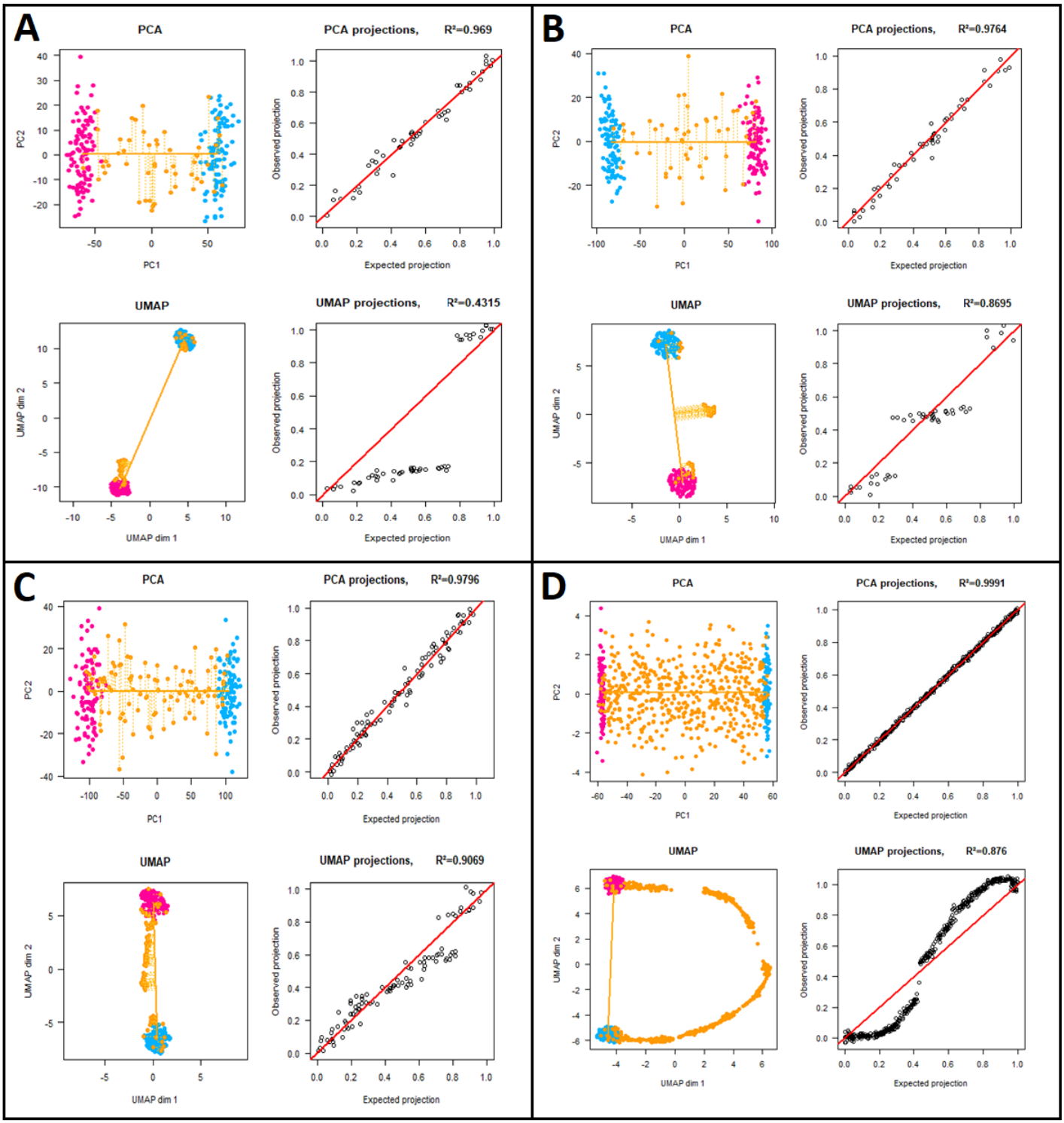

Figure 7 : Comportements de UMAP, réduction de dimension de données simulées avec l'ACP et UMAP, avec le graphique des projections observées en fonction des projections théoriques

En bleu et violet sont représentées les échantillons des deux groupes de cancers, et en orange les intermédiaires. **A** : « L'effet gravité » de UMAP groupe les échantillons intermédiaires dans les clusters de cancer. **B** : « L'effet gravité » de UMAP forme un cluster intermédiaire. **C** : « L'effet gravité » de UMAP rend discontinue la continuité des intermédiaires. **D** : UMAP représente le continuum d'intermédiaires sur un arc de cercle lorsque le nombre d'échantillons intermédiaire est élevée

Discussion

L'ACP pour identifier les cancers phénotype intermédiaire

Dans le cadre de notre modèle, où nous avons deux groupes de cancers avec des intermédiaires, nous avons montré que l'ACP est une très bonne méthode de réduction de dimension pour représenter les échantillons de phénotypes intermédiaires. Ceci est principalement du au fait que la première CP est associé à la variance entre les deux groupes de cancers. L'hypothèse de linéarité de l'ACP permet alors de correctement placer les intermédiaires puisque selon le modèle, les intermédiaires sont le résultat d'une combinaison linéaire des profils génomique des deux groupes. Dans le cas d'un autre modèle où la variance entre les groupes de cancers ne pourrait pas être associé à une ou deux CP, par exemple lorsque l'on s'intéresse à plus de deux groupes de cancers, l'ACP aurait une piètre performance dans la représentation des intermédiaires dans un espace à deux dimensions. Théoriquement, pour représenter correctement la variation entre K groupes, K-1 CP seraient nécessaires (Hastie T *et al.,* 2009). Lorsque K est grand, la visualisation de la totalité de la variation deviendrait difficile. Dans ce cas, UMAP pourrait être une alternative.

UMAP pour identifier les cancers phénotype intermédiaire

UMAP est une méthode de réduction de dimension qui devient populaire pour l'analyse de données génomique, cependant comme nous l'avons montré, il existe des situations où cette méthode ne permet pas de représenter les phénotypes intermédiaires, et ceux-ci sont groupés avec d'autres cancers. Cela peut être problématique pour le diagnostic et la prise en charge des maladies rares.

Dans le cadre de notre modèle, UMAP représente généralement moins bien les intermédiaires que l'ACP. UMAP permet tout de même de correctement représenter les échantillons de phénotype intermédiaire quand le nombre d'échantillons et la proportion d'intermédiaires sont grands, et quand les groupes de cancers sont moléculairement ni trop proches, ni trop éloignés. Autrement, UMAP ne permet pas une bonne visualisation du continuum entre les deux cancers. Augmenter la valeur du paramètre *min_dist* permet de mieux faire ressortir les intermédiaires « cachées » dans les données avec des valeurs de *min_dist* plus faible. Un autre paramètre de UMAP pouvant avoir un effet sur l'indentification des intermédiaires, qui reste à être exploré, est le paramètre *n_neighbours*.

Discussion sur la méthode d'évaluation

Le coefficient de détermination (R^2) des projections des intermédiaires est une bonne métrique pour faire une évaluation rapide et simple des méthodes de réduction de dimension. Cependant il est possible que des bonnes représentations de données, c'est-à-dire avec une bonne continuité d'intermédiaires entre les groupes, produisent des R^2 faibles, ou inversement que des mauvaises représentations produisent de bons R^2. Ceci est en particulier le cas avec UMAP lorsque la continuité des intermédiaires n'est pas représentée par un segment mais par une courbe.

Il serait alors intéressant de proposer une autre métrique prenant en compte ce phénomène pour avoir une évaluation plus juste de la méthode UMAP, en utilisant des interpolations par exemple comme le fit d'une spline non-linéaire.

Discussion sur le modèle

Le modèle utilisé dans cette étude est proche de la description des modèles de tumeurs biphasiques, où la tumeur est une mixture de cellules de types et tous les gènes différentiellement exprimés des intermédiaires ont une valeur intermédiaire entre les deux types (Timothy A. Yap *et al.*, 2017). Dans le cadre de l'identification de phénotype intermédiaires, un autre modèle possible est celui décrivant une tumeur en cours de transformation, où les gènes ne sont pas tous différentiellement exprimées simultanément. Hypothétiquement, cet autre modèle ne devrait pas générer des résultats différents de ceux obtenus dans cette étude sur l'évaluation des méthodes de réduction de dimension.

Dans notre modèle le coefficient de mixture des échantillons intermédiaires est distribué uniformément. En réalité il existe des cas où les deux clusters de cancer correspondent à des états stables et les phénotypes intermédiaires à mi-chemin entre ces deux états sont très rares. Pour modéliser ce phénomène une distribution beta peut être utilisée.

Perspectives et Conclusion

Nous avons proposé un modèle et une méthode pour évaluer la capacité des algorithmes de réduction de dimension à correctement représenter des cancers de phénotype intermédiaire. Nos résultats ont montré que, dans le contexte où l'on cherche à identifier des cancers de phénotype intermédiaire entre deux groupes de cancers, l'ACP un algorithme de réduction de dimension efficace à cette tâche. UMAP en revanche est un algorithme qui dans ce contexte à une efficacité moindre dans la représentation des cancers de phénotype intermédiaires, notamment lorsque le nombre d'échantillon est faible.

La suite logique serait de proposer une méthode de clustering pour identifier des cancers de phénotype intermédiaire dans un jeu de données dont la réalité n'est pas connue. Il serait alors possible avec le modèle d'évaluer cette méthode de clustering en examinant les taux de vrai positifs, faux positifs, vrai négatifs et faux négatifs. De plus, de la même manière que l'exploration des paramètres dans cette étude, il serait intéressant d'explorer les paramètres influant la méthode de clustering pour découvrir quels réglages, à la fois de la méthode de clustering et de la méthode réduction de dimension, sont optimaux pour l'identification de cancer de phénotype intermédiaires.

Références

Alex Diaz-Papkovich *et al.* Revealing multi-scale population structure in large cohorts, 2019. *bioRxiv 423632*

Etienne Becht *et al*. Dimensionality reduction for visualizing single-cell data using UMAP, 2019. *Nature Biotechnology 37,38-44*

Hastie T *et al.* The Elements of Learning: Data Mining, Inference, and Prediction, 2009. *Springer, New York*

Jake Lever *et al.* Principal component analysis, 2017. *Nature Methods 14, 641-642*

Katherine A. Hoedley *et al.* Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer, 2018. *Cell 173-2*

Lan T. Jolliffe *et al.* Principal component analysis : a review and recent developments, 2016. *Philos Trans A Math Phys Eng Sci 374(2065)*

Leland McInnes *et al.* UMAP : Uniform Manifold Approximation and Projection for Dimension Reduction, 2018. *arXiv :1802.03426*

Nicolas Alcala et *al.* Redefining mesothelioma types as a continuum uncovers the immune and vascular systems as key players in the diagnosis and prognosis of this disease, 2018. *bioRxiv 334326*

Pranav Nerurkar *et al.* Empirical Analysis of Data Clustering Algorithms, 2018. *Procedia Computer Science 125, 770-779*

Timothy A. Yap *et al*. Novel insights into mesothelioma biology and implications for therapy, 2017. *Nature Review Cancer 17, 475-488*