



Université Claude Bernard



Lyon 1

## Master 2 Bio-informatique Moléculaire : Méthodes et Analyses

2024-2025

Opérée au sein de :  
**Université Claude Bernard Lyon 1**

**UE** : UE-BIO2461M Stage Entreprise / Laboratoire 2

**Stage** : Centre de Recherche en Cancérologie de Lyon  
Équipe Saintigny – Analyse intégrée de la dynamique du cancer

**Encadrant** : Dr Pierre Martinez

## Bio-informatique et cancer : TITRE DE OUf

---

**Jordan Dutel**

## Résumé

Ce stage réalisé dans l'équipe "Analyse intégrée de la dynamique du cancer" dirigé par Dr Pierre Saintigny, et encadré par Dr Pierre Martinez s'inscrit dans une initiative nationale pour améliorer la prise en charge des patientes atteintes de formes rares et graves de cancer du sein, appelé cancer du sein métaplasique (MpBC). Ce projet implique plusieurs centres de recherche tels que le Centre de Recherche en Cancérologie de Lyon et le Centre Léon Bérard (Lyon), l'Institut Curie (Paris), l'Institut Bergonié (Bordeaux) et le Centre de Recherche en Cancérologie et Immunologie (Nantes). Ce travail de recherche vise à combler nos connaissances encore incomplètes sur la biologie des MpBC. En effet, la prise en charge de ces tumeurs est confrontée à des lacunes pour le diagnostic et à un manque important d'options thérapeutiques. Par des méthodes de transcriptomique spatiale et d'analyses des altérations du nombre de copies (CNA), j'ai réussi à identifier certains marqueurs phénotypiques d'intérêt pour caractériser les sous-types de ces carcinomes. De plus, j'ai également pu mettre en évidence des régions chromosomiques significativement altérés pour des sous-types tumoraux spécifiques. Ces résultats pourront contribuer à des analyses approfondies qui seront réalisées prochainement, et ouvriront la voie vers l'identification de marqueurs pour le diagnostic et de cibles thérapeutiques pour la prise en charge des patientes atteintes de ces cancers du sein.

# Table des matières

<b>Liste des abréviations</b>	<b>4</b>
<b>Liste des logiciels utilisés</b>	<b>5</b>
<b>1 Introduction</b>	<b>6</b>
1.1 Contexte et état de l'art . . . . .	6
1.2 Présentation des MpBC . . . . .	6
1.3 Problématique(s) . . . . .	7
1.4 Pertinence de la transcriptomique spatiale . . . . .	8
<b>2 Matériels et méthodes</b>	<b>9</b>
2.1 Echantillons MpBC . . . . .	9
2.1.1 Cohorte : CLB . . . . .	9
2.1.2 FFPE fixation et H&E coloration . . . . .	9
2.1.3 Séquençage Visium & alignement . . . . .	9
2.2 Annotation phénotypique des spots . . . . .	10
2.3 Données scRNA-seq . . . . .	10
2.3.1 Contrôle qualité et filtrage des spots . . . . .	10
2.3.2 Normalisation et Scaling . . . . .	10
2.4 Analyse des marqueurs phénotypiques . . . . .	11
2.4.1 Harmony . . . . .	11
2.4.2 Seurat . . . . .	11
2.5 Analyse des altérations génétiques . . . . .	12
2.5.1 InferCNVPlus . . . . .	12
2.5.2 Définition des altérations génomiques divergentes . . . . .	14
2.6 Software et packages . . . . .	14
2.6.1 RStudio et language R . . . . .	14
<b>3 Résultats</b>	<b>15</b>
3.1 Analyse des marqueurs phénotypiques . . . . .	15
3.1.1 Réduction de dimensionnalité . . . . .	15
3.1.2 Clustering et enrichissement archétypal . . . . .	15
3.1.3 Identification des gènes marqueurs . . . . .	17
3.2 Analyse des altérations génotypiques divergentes . . . . .	19
3.2.1 Profils génomiques biphasiques . . . . .	19
3.2.2 Identification des altérations divergentes entre compartiments pairés . . . . .	20
3.2.3 Visualisation des différences de score CNA entre compartiment . . . . .	21
<b>4 Discussion</b>	<b>25</b>
4.1 Limites de la technologie Visium (sous-titres à supprimer) . . . . .	25
4.1.1 Risque d'hétérogénéité cellulaire intra-spot . . . . .	25
4.2 Perspectives du projet (sous-titres à supprimer) . . . . .	28

4.2.1	Consortium MAESTRO : échantillons additionnels . . . . .	28
4.2.2	Analyse de réseaux / Pathways dérugglés plutôt que des gènes spécifiques .	29
4.2.3	Analyse approfondie via snRNA-seq . . . . .	29
4.2.4	Microdissection des compartiments . . . . .	30
4.2.5	Déterminants non génétiques . . . . .	31
4.2.6	Validation IHC & Analyses fonctionnelles . . . . .	31
<b>5</b>	<b>Conclusions</b>	<b>32</b>

## Table des figures

1	Exemple d'échantillon MpBC mixte . . . . .	7
2	Correction du batch effect entre les 16 échantillons distincts de MpBC . . . . .	16
3	Projection UMAP des données transcriptomiques Visium . . . . .	17
4	Visualisation de l'expression génique par cluster sur la projection UMAP . . . . .	18
5	Profil d'expression des gènes marqueurs sélectionnés selon les types cellulaires tumoraux et les patients atteints de MpBC . . . . .	19
6	Représentation des scores CNA obtenus avec InferCNVPlus, avant (A) et après normalisation (B), en fonction des cytobandes mineures pour chaque chromosome du génome . . . . .	21
7	Profil de significativité des altérations chromosomiques pour le patient MpBC9 . .	22
8	Carte de chaleur des altérations du nombre de copies pour le patient MpBC9 . .	23
9	Analyse de la corrélation des scores CNA entre les deux sous-types tumoraux du patient MpBC9 . . . . .	24
10	Taille de l'effet des altérations chromosomiques entre sous-types tumoraux . . . . .	24

## Liste des tableaux

1	Tableau récapitulatif des différents tissus (sous-type tumoral) comparés dans l'analyse des CNA par patient . . . . .	13
---	-----------------------------------------------------------------------------------------------------------------------	----

## Liste des abréviations

- TNBC** : Triple Negative Breast Cancer
- MpBC** : Metaplastic Breast Carcinoma
- CNA** : Copy Number Alteration
- CNV** : Copy Number Variant
- CLB** : Centre Léon Bérard
- H&E** : Hematoxylin & Eosin
- FFPE** : Formalin-Fixed, Paraffin-Embedded
- CNRS** : Centre National de la Recherche Scientifique
- UMI** : Unique Molecular Identifier
- UMAP** : Uniform Manifold Approximation and Projection
- PCA** : Principal Component Analysis
- DGEA** : Differential Gene Expression Analysis
- KNN** : k-Nearest Neighbors
- ID** : Identifiant
- UCSC** : University of California Santa Cruz
- MAST** : Model-based Analysis of Single cell Transcriptomics
- snRNA-seq** : single nuclei RNA-sequencing
- RCTD** : Robust Cell Type Decomposition
- SCENIC** : Single-Cell rEgulatory Network Inference and Clustering
- MAESTRO** : MetaplAstic brEaST caReinOma
- IHC** : Immuno-Histo-Chimie

## Liste des logiciels utilisés

**RStudio** : (version 2024.09.0+375 « Cranberry Hibiscus »)

**Seurat** : (v5.1.0)

**Harmony** : (v1.2.3)

**MAST** : (v1.32.0)

**InferCNVPlus** : (v3.20)

**Loupe Browser** : (v8.1.2, 18 Nov. 2024)

**Space Ranger** : (v2.0.0)

# 1 Introduction

## 1.1 Contexte et état de l'art

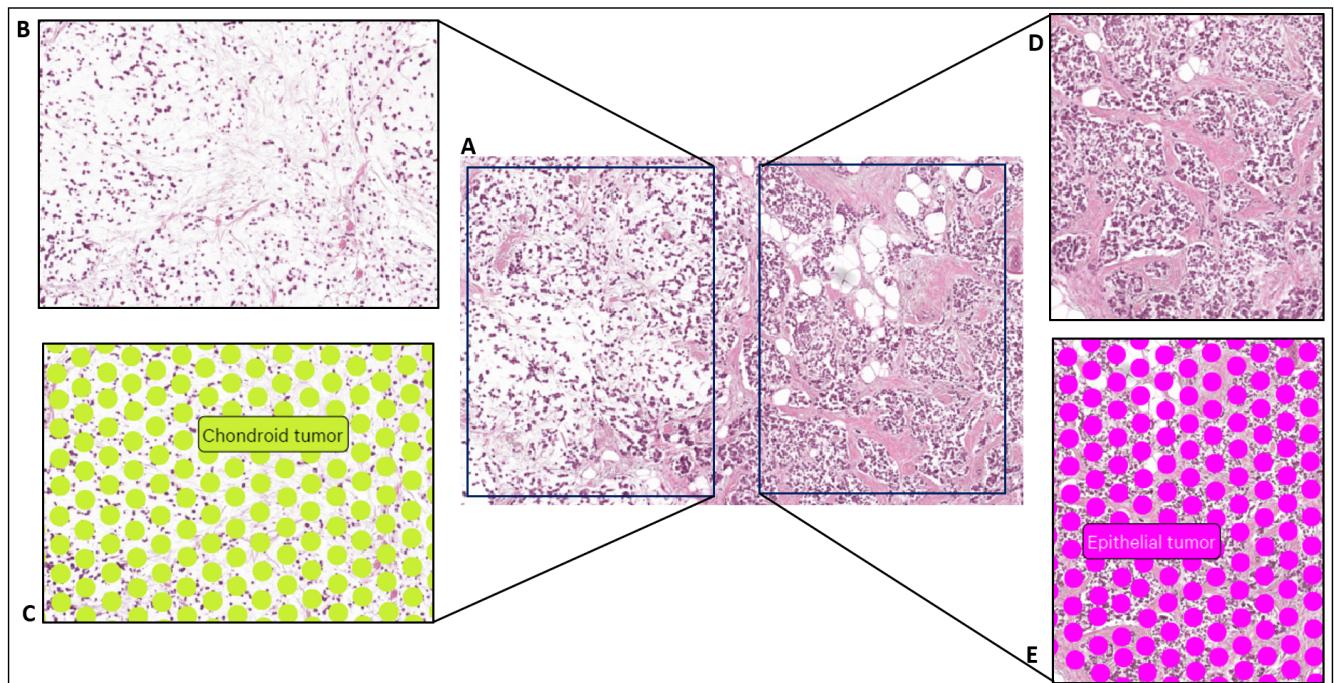
Selon l'OMS (Organisation Mondiale de la Santé), le cancer du sein était la première cause de cancer chez les femmes dans 157 pays sur 185 [1] en 2022, et on considère qu'environ 1 femme sur 12 en sera diagnostiquée au cours de sa vie [1]. Aucun facteur de risque spécifique autre que le sexe et l'âge n'est encore connu [1] [2], faisant donc de cette pathologie un enjeu majeur de santé publique dans les années à venir. Parmi les cancers du sein, on distingue les types TNBC (Triple Negative Breast Cancer), caractérisés par l'absence de certains types de récepteurs aux œstrogènes et progestérone. Les plus connus de ces récepteurs sont les protéines HER2 (Human Epidermal growth factor Receptor 2) [3] et BRCA1/2 (Breast Cancer 1/2) [4]. De plus, l'absence de ces récepteurs pour les patientes atteintes de TNBC rend le traitement par des thérapies ciblées extrêmement compliqué. Ainsi, les TNBC sont la plupart du temps des tumeurs agressives et dites "de haut grade", c'est-à-dire, qu'elles présentent d'importantes modifications cellulaires dysplasiques (dont le phénotype est très différent de celui des cellules normales). Elles disposent également d'une capacité à se développer et à se propager très rapidement [5] [6]. Enfin, ces TNBC sont caractérisés par une variété impressionnante de sous-types, avec des phénotypes très hétérogènes [7]. Parmi eux, nous retrouvons des sous-types rares, tels que les Basal-Like (Basal-Like Breast Cancer, BLBC) et les carcinomes du sein métaplasiques (Metaplastic Breast Cancer, MpBC) [8]. Ces derniers sont des cas rares et complexes de TNBC, encore aujourd'hui très mal compris et avec aucun marqueur moléculaire pour le diagnostic. Les patientes sont donc confrontées aujourd'hui à un manque important d'options thérapeutiques, ce qui en fait une forme de cancer très aggressive et avec une forte mortalité.

## 1.2 Présentation des MpBC

Dans ce projet de recherche nous nous intéressons plus spécifiquement aux MpBC car les cellules composant ce sous-type disposent d'une capacité remarquable à se trans-différencier. Selon le CNRS, la transdifférenciation est « la conversion d'un type cellulaire entièrement différencié en un autre type » [9]. En effet, les MpBC sont caractérisés par une importante plasticité cellulaire aboutissant à des compartiments tissulaires phénotypiquement très distincts, au sein même de la tumeur. Les échantillons MpBC présentant au moins 2 compartiments cellulaires au sein de la tumeur sont dits « mixtes », en opposition aux MpBC « pures », ne présentant pas encore de cellules cancéreuses trans-différenciées.

Actuellement, on dénombre 5 types de trans-différenciation des MpBC à partir des cellules tumorales épithéliales : Malpighienne (ou Squameuse), Mésenchymateuse, Spindle-like (ou Fusiforme), Chondroïde et Osteosarcomatoïde. Selon le type de la trans-différenciation, les cellules tumorales peuvent par exemple présenter un phénotype plus épithelial que les cellules cancéreuses initiales (trans-différenciation malpighienne), ou au contraire présenter un phénotype plus stromal (trans-différenciation mésenchymateuse). Certaines cellules, lors de ce processus peuvent également prendre une forme allongée et fusiforme, en forme d'épingle (trans-différenciation Spindle-like), ou bien créer abondamment de la matrice extracellulaire autour d'elles et s'enfermer dans une logette cartilagineuse (trans-différenciation chondroïde), voire même une matrice

ostéoïde ressemblant à de l'os immature (trans-différenciation ostéosarcomatoïde). La **figure 1** représente un exemple de différents compartiments tumoraux retrouvés au sein d'un échantillon MpBC mixte. Cette large diversité de différenciation des MpBC n'est, encore aujourd'hui, pas totalement comprise et les mécanismes moléculaires restent inexpliqués. C'est à ce jour une lacune importante dans notre compréhension de la plasticité des cancers. Or, en appréhendant l'origine de ces mécanismes, nous serions plus à même de développer des options de diagnostic moléculaire précis et d'élargir notre panel de cibles thérapeutiques pour ces cancers agressifs, qui pour l'instant n'ont aucun traitement disponible.



**Figure 1 :** Exemple d'échantillon MpBC mixte. Au sein d'une même tumeur (A), une zone de carcinome mammaire chondroïde invasif est visible à gauche (B) avec des spots Visium superposé sur le tissu (C), tandis qu'une zone de transdifférenciation épithéliale est observable à droite (D), avec les spot Visium (E).

FIGURE 1 – Exemple d'échantillon MpBC mixte

### 1.3 Problématique(s)

Dans ce projet de recherche, afin de comprendre l'origine des mécanismes sous-tendant l'apparition de ces différents compartiments cellulaires au sein des MpBC, mon travail s'est articulé autour de 2 principaux objectifs. Dans un premier temps, grâce à des données de transcriptomique spatiale pour chaque tissu, je me suis intéressé au profil transcriptionnel de chaque sous-type cellulaire au sein des MpBC, afin de trouver des marqueurs phénotypiques spécifiques. Dans un second temps, je me suis focalisé sur les possibles causes génétiques générant ces différents sous-types, en effectuant une analyse des altérations du nombre de copie (CNA) de chaque compartiment tumoral.

## 1.4 Pertinence de la transcriptomique spatiale

Afin de répondre précisément à ces questions, nous avons basé nos analyses sur des données de transcriptomique spatiale générées pour 16 échantillons MpBC de patientes. Cette technologie à résolution spatiale est particulièrement intéressante dans ce projet de recherche car elle va permettre d'associer des annotations pathologiques des sous-types cellulaires, réalisées par une experte anatomopathologiste, avec des analyses multi-omiques. La nature même des MpBC fait qu'il existe plusieurs compartiments cellulaires au sein même de la tumeur. L'utilisation de technologie seule, tel que le bulk-RNAseq, serait non pertinente dans ce contexte, car nous perdrons l'information spatiale et celle du type cellulaire séquencé. Ainsi, avec une information spatiale en plus du profil transcriptomique, il sera possible d'établir des marqueurs histologiques et moléculaires spécifiques à chaque sous-type cellulaire trans-différencié, qui pourront être utilisés par la suite pour le diagnostic et l'identification de potentielles nouvelles thérapies.

## 2 Matériels et méthodes

### 2.1 Echantillons MpBC

#### 2.1.1 Cohorte : CLB

Les échantillons MpBC proviennent d'une cohorte de 39 patientes ayant eu un diagnostic de carcinome mammaire métaplastique et ayant suivies au Centre Léon Bérard (CLB). 16 patientes présentant des MpBC mixtes, ou biphasiques, avec des transdifférenciations de différents types, ont été sélectionnées pour analyse approfondie. La participation au projet de recherche s'est faite via le recueil de la non opposition des patientes, selon le cadre éthique mis en place par le CLB. Après inspection a posteriori, deux échantillons n'ont pas permis d'obtenir 2 types de compartiments tumoraux différents par coupe (MpBC6 et MpBC7)

#### 2.1.2 FFPE fixation et H&E coloration

Une fois les prélèvements cliniques réalisés pour le diagnostic, les échantillons réséqués ont suivi un protocole FFPE (Formalin-Fixed, Paraffin-Embedded) de fixation et d'inclusion en paraffine, afin de les conserver durablement et de faciliter leur réutilisation à des fins cliniques ou de recherche ultérieure. La gestion des échantillons et la réalisation des protocoles ont été assuré par la Plateforme de Gestion des Echantillons Biologiques (PGEB) du CLB [10]. Pour chaque patiente, une partie du tissu a été découpé au microtome puis placé sur une lame histologique afin de réaliser une coloration H&E (Hématoxilin et Eosine). Le reste de la tumeur a été conservé à température ambiante (20°C) jusqu'à réutilisation.

#### 2.1.3 Séquençage Visium & alignement

Les tissus prélevés ont été ensuite placés sur lame de séquençage Visium « Human Transcriptome Probe Panel v2 » de la société 10X Genomics. Cette technologie, appropriée pour les échantillons FFPE, permettant de combiner information spatiale et transcriptomique, via des spots de cellules (1 spot représentant environ 10 cellules, 55µm environ), qui permettent de déterminer à quel endroit de la coupe un transcrit unique est exprimé. Une lame de séquençage Visium contient 2 surfaces de capture et chaque surface est segmenté en 4992 spots différents recouvrant l'entièreté de l'échantillon. Cette technologie est une méthode UMI-based permettant la quantification d'ARN à partir d'une molécule originale marquée par un identifiant unique (UMI), ou barcode, puis une amplification de ces ARN par PCR [11]. Une surface de capture peut capturer jusqu'à 18000 gènes. Pour notre projet, le séquençage s'est fait avec une profondeur moyenne de 60k reads par spot pour le batch de séquençage 1 (MpBC1 à 8), et 25k reads par spot pour le batch de séquençage 2 (MpBC9 à 16). Parmi les 16 échantillons, 1 seul présente des statistiques de séquençage et de contrôle qualité non satisfaisantes (MpBC12), nous l'avons donc exclu des analyses.

Pour chaque échantillon séquencé, les *reads* ont ensuite été démultiplexés et alignés contre le génome humain (*version GRCh38*), via un pipeline entièrement automatisé géré par la plateforme bio-informatique Gilles Thomas du CLB, via le logiciel Space Ranger (version 2.0.0)

[12].

## 2.2 Annotation phénotypique des spots

Afin de maximiser l'information spatiale, nous avons annoté chacun des spots de chaque échantillon en catégorisant les sous-types cellulaires observé comme majoritaire sur la lame pairee avec coloration H&E. Pour cela nous avons combiné les informations du sous-type cellulaire fournies par 2 méthodes différentes.

Une première analyse de clustering des spots a été réalisé par le logiciel « Loupe Browser » (v8.1.2) [13] de la société 10X Genomics, permettant de regrouper les spots présentant des profils transcriptomiques similaires par coupe, via un clustering k-means.

Puis dans un deuxième temps, les clusters ont été individuellement vérifiés et manuellement annotés par Dr Isabelle Treilleux, expert anatomopathologiste au sein du CLB. Le nombre de clusters par coupe a également été déterminé selon la concordance avec les observations histologiques. Cela nous garantit donc une bonne confiance quant à l'identité des spots de cellules séquencés sur la lame, pour chaque patiente.

## 2.3 Données scRNA-seq

### 2.3.1 Contrôle qualité et filtrage des spots

Avant toute analyse approfondie des données de transcriptomique spatiale, un contrôle qualité a été réalisé. Dans un premier temps, pour chaque échantillon MpBC, la matrice de comptes a été filtré selon certaines caractéristiques : le nombre d'UMI (Unique Molecular Identifier) compté par spot de cellule, et le nombre de gène différents compté par spot. Dans notre analyse, nous avons sélectionné uniquement les spots de cellule avec un nombre d'UMI et un nombre de features (gènes) supérieur à 500. Cela permet d'éliminer les spots de mauvaise qualité et/ou avec trop peu d'activité transcriptionnelle détectée, probablement liés à des problèmes de séquençage et des cellules endommagées. Une limite supérieur correspondante au 99e percentile de l'échantillon a été appliquée pour retirer les spots anormalement actifs.

### 2.3.2 Normalisation et Scaling

**Nombre de gènes sélectionnés** La normalisation des matrices de comptage s'est faite après la concaténation des différentes matrices individuelles pour chaque patiente (au préalablement filtrées pour enlever les spots de mauvaise qualité) en une seule matrice globale, regroupant tous les échantillons. Pour cela les comptes d'UMI pour chaque spot de cellule ont été log-normalisés avec la fonction LogNormalize() du package Seurat (version v5.1.0) [14]. Cette log-normalisation permet à la fois de compenser les potentielles différences de profondeur de séquençage entre les différents spots et les différents échantillons, mais également à stabiliser la variance des comptes UMI en réduisant l'impact des spots avec des comptes UMI extrêmes.

De plus, pour les données transcriptomiques brutes de chaque spot, on peut considérer que certains gènes sont moins informatifs que d'autres, et qu'ils contribueront plus à amener un bruit de fond dans les analyses que plutôt une réelle information biologique. Nous avons donc réalisé

une seconde étape correspondant à la sélection des gènes les plus variables du dataset, nous permettant ainsi de réduire une première fois la complexité du jeu de données, tout en conservant la majorité du signal biologique. Pour cela, parmi tous les gènes exprimés dans le jeu de données, nous avons sélectionné les 750 gènes les plus variables. Ce seuil a été déterminer de façon empirique en fonction de la qualité des résultats obtenues en aval. Aujourd’hui, la détermination de ce seuil fait l’objet de nombreux débats au sein de la communauté scientifique et plusieurs études scientifiques tentent de montrer un seuil optimal. Cependant il n’existe actuellement que des plages de valeurs recommandé, généralement entre 500 et 2000 gènes, mais la détermination exacte de ce seuil est très dépendante de l’outil d’analyse utilisé et propre au dataset, et le choix final est souvent laissé à la discréption de l’utilisateur. Pour notre jeu de donnée, la sélection de 750 gènes permet de réduire significativement la dimensionnalité et donc la complexité, mais également d’éviter d’intégrer dans notre analyse des gènes peu variables, c’est-à-dire peu informatifs, correspondant à du bruit. Toutefois, ce seuil de gène les plus variables reste suffisant pour permettre de capturer l’essentiel de l’information biologique contenu dans nos données.

Enfin, une dernière étape correspondant au scaling, nous a permis de transformer les données d’expressions transcriptomique par gène, pour que chacun d’eux aient une moyenne centrée en 0 et une variance réduite à 1. En effet, même avec une log-normalisation appliquée à l’étape précédente, certains gènes sont exprimés à des échelles très différentes. Leur grande variance peut alors fausser les réductions de dimension réalisée par la suite. Avec cette étape, on équilibre donc la contribution de chaque gène.

## 2.4 Analyse des marqueurs phénotypiques

### 2.4.1 Harmony

**Correction batch-effect** Notre analyse des marqueurs phénotypiques intègre 16 patientes différentes et donc autant d’échantillons MpBC. Nous avons donc corrigé l’effet batch de notre jeu de données en utilisant le package R « Harmony » (version v1.2.3) [15]. Après une réduction PCA de notre matrice de comptage globale sur 50 composantes, les variations non biologiques ont été contrôlé avec la fonction `RunHarmony()` en spécifiant toutes les co-variables pouvant biaiser l’analyse. Nous avons spécifié les métadonnées dont les effets indésirables sont à corriger et, dans notre cas, correspondantes à l’ID de la patiente, le lot de séquençage et l’ID de la lame Visium utilisé pour le séquençage (2 patients différents par lame). Les autres paramètres ont été déterminé de façon empirique et selon les conseils de la communauté scientifique. Ainsi, pour éviter une sur-correction, nous avons appliqué une pénalité spécifique à chaque variable (`theta`) égale à 2. Enfin, nous avons précisé un `lambda = 1`, un `sigma = 0.2` et `nclust = 150`.

### 2.4.2 Seurat

**UMAP** La détermination des paramètres pour réaliser la projection UMAP (Uniform Manifold Approximation and Projection), afin de réduire la dimensionnalité des données et ainsi aider à leur analyse et interprétation, s’est faite de façon empirique. Dans notre cas, nous avons utilisé comme base de projection les 20 premières composantes de l’espace de dimension ré-

duit normalisé par le package Harmony, avec un nombre de voisin (`n.neighbors`) de 75, une distance minimale entre les points de l'espace (`mindist`) de  $10^{-4}$  et dispersion globale des points (`spread`) de 2. Enfin, nous avons favorisé la connectivité locale (`set.op.mix.ratio = 1`) et utilisé une métrique « cosine » pour mesurer la distance entre les cellules sans l'espace de dimension réduit.

**Clustering** Afin de réaliser le clustering des spots de cellules présentant des profils transcriptomiques similaires, nous avons tout d'abord construit le graphe KNN (k-Nearest Neighbors) en utilisant les composantes principales corrigées par Harmony comme base du graphe. Pour cela nous avons effectué la construction du graphe sur les 20 premiers axes de la réduction Harmony. Puis nous avons appliqué un algorithme de clustering de type Louvain. Le paramètre contrôlant la granularité du clustering (`resolution`), a été, là aussi, déterminé de façon empirique et fixé à 0.15. Pour déterminer la valeur de résolution optimale, nous avons notamment regarder la stabilité des clusters, le nombre de types cellulaires attendues (correspondant aux différents sous-types tumoraux) et en comparant les compositions en spot des clusters formés selon les types cellulaires annotés par l'experte anatomopathologiste.

**Expression différentielle** Une analyse des gènes différentiellement exprimés (DGEA) a été réalisé pour chaque cluster afin de déterminer les cibles moléculaires d'intérêt pour chaque sous-type tumoral. Pour cela, nous avons utilisé la fonction `FindAllMarkers()` de Seurat, avec l'option MAST (Model-based Analysis of Single cell Transcriptomics) comme test pour identifier, parmi tous les gènes, ceux étant les plus différentiellement exprimés. Cette méthode est plus adaptée au données UMI-based, qui peuvent créer de nombreuses égalités de classement préjudiciables lorsque la méthode standard de test de Wilcox est utilisée. De plus, afin de filtrer les résultats, nous avons nous sommes focalisés sur les gènes présentant un logFC (log Fold-Change) supérieur à 2. Dans notre projet, nous ne récupérons que les gènes sur-exprimés dans les compartiments transdifférenciés, car nous cherchons des cibles spécifiques permettant de les identifier de manière moléculaire pour le diagnostic. Enfin, un dernier filtre a été appliqué afin de ne conserver que les marqueurs qui sont exprimé dans au moins 50% de la population des spots/cellules du cluster. Cela permet d'ôter les marqueurs trop spécifiques à un sous-groupe du cluster et donc non représentatif du sous-type tumoral en général. Cela est particulièrement pertinent lorsqu'il y a de fortes variations de taille (i.e. nombre de spots) entre les différents échantillons pour un sous-type donné.

## 2.5 Analyse des altérations génétiques

### 2.5.1 InferCNVPlus

**Constitution du groupe de cellule de références** InferCNVplus [16], est une extension de la méthode inferCNV [17], et permet d'inférer les CNV les plus probables dans un jeu de données transcriptomiques, en comparant l'expression des gènes dans les cellules tumorales par rapport à celle de cellules normales de référence. Il va également prendre en compte la position des gènes dans le génome et identifier les variations d'expression des régions chromosomiques correspondant à un profil d'altération CNV. A chaque région, l'outil attribue un score de CNV

entre -1 (délétion) et 1 (amplification).

L'analyse des altérations génétiques de chaque sous-type des échantillons MpBC, s'est faite à l'aide du package R « InferCNVplus » (version v3.20). Pour cela un groupe de spots de référence a été construit à partir des spots annotés comme non tumoraux dans chaque échantillon MpBC. Comme chaque échantillons MpBC ne contenaient pas systématiquement des spots de cellules annotés comme non tumoraux, nous avons créé un groupe de référence unique pan-échantillon, en combinant tous les spots de tissu normal non tumoral. Ainsi, lors de l'analyse des altérations génétiques des compartiments tumoraux, chaque échantillons MpBC ont été comparé à un groupe de référence, commun à tous les échantillons. Les spots considérés comme non tumoraux étaient annotés comme contenants des cellules sanguines, des cellules épithéliales normales et des cellules mésenchymateuses normales.

**Profils génomiques par cytobande** Les scores CNA par spot fournis par InferCNVPlus nous ont permis d'établir un profil des altérations génomiques pour chaque compartiment trans-différencié de chaque échantillon MpBC. Pour cela, nous avons utilisé la segmentation en cytobande mineure du génome humain (version GRCh38 de l'UCSC [18]). Chaque gène présent dans la matrice de comptage initiale a été attribué à une des cytobandes mineures du génome humain, selon ses coordonnées. Les scores CNA ont été agrégés en faisant la médiane des scores de tous les gènes appartenant à chaque cytobande mineure, pour chaque patient, et chaque sous-type tumoral. Ainsi, pour chaque cytobande mineure de chaque patient, nous avons obtenu un score CNA médian pour chacun des 2 types de compartiment tumoral considéré dans l'échantillon MpBC, le long de tout le génome humain. Les échantillons MpBC purs (MpBC6 et 7), ne contenant par définition qu'un seul type de compartiment tumoral, ont été exclus de l'analyse. La **table 1** récapitule les comparaisons de score CNA des tissus tumoraux par patient.

Patients	Sous-type tumoral 1	Sous-type tumoral 2
MpBC1	Épithérial	Mésenchymateux
MpBC2	Épithérial	Fusiforme (Spindle-like)
MpBC3	Malpighien	Fusiforme (Spindle-like)
MpBC4	Ostéosarcomatoïde	Fusiforme (Spindle-like)
MpBC5	Épithérial	Fusiforme (Spindle-like)
MpBC8	Épithérial	Mésenchymateux
MpBC9	Épithérial	Chondroïde
MpBC10	Épithérial	Fusiforme (Spindle-like)
MpBC11	Épithérial	Fusiforme (Spindle-like)
MpBC13	Épithérial	Chondroïde
MpBC14	Épithérial	Fusiforme (Spindle-like)
MpBC15	Épithérial	Chondroïde
MpBC16	Épithérial	Chondroïde

TABLE 1 – Tableau récapitulatif des différents tissus (sous-type tumoral) comparés dans l'analyse des CNA par patient

**Normalisation** Les spots de la technologie Visium correspondent à environ  $55\mu\text{m}$  de diamètre et peuvent donc contenir plusieurs cellules tumorales et non tumorales. Par conséquent, la présence dans les spots, annotés tumoraux, de cellules non tumorales peut biaiser ces scores CNA, car les cellules non tumorales ne présentent pas d'altération du nombre de copies. Les compartiments tumoraux de différents sous-types peuvent ainsi présenter des variations à la fois en densité de cellule (plus de cellules par spots correspondant à plus de transcrits), et/ou d'infiltration par des cellules non tumorales qui peuvent impacter les scores CNA obtenus par inferCNV. Afin de normaliser ces scores CNA entre les différents compartiments tumoraux pour chaque patiente, nous avons donc factorisé chaque score CNA par un facteur d'échelle, permettant de faire coïncider les extrêmes des scores de délétion ( $\text{score CNA} < 0$ ) et d'amplification ( $\text{score CNA} > 0$ ) entre les paires de compartiments tumoraux de chaque patiente. Un facteur d'amplification a été défini comme le rapport entre le score CNA maximal du tissu 1 et celui du tissu 2. De même, un facteur de délétion a été calculé en prenant la valeur absolue du rapport entre les scores CNA minimaux des deux tissus. Dans chaque cas, seul le tissu ne présentant pas l'extrême global (maximum ou minimum, tous tissus confondus) a vu ses scores CNA multipliés par ce facteur, afin d'harmoniser l'échelle des valeurs entre les tissus. Ainsi, les biais liés à la profondeur de séquençage ne perturbent plus l'analyse. Un spot contenant peu de cellules et donc une faible profondeur de séquençage (par exemple le sous-type chondroïde) est ainsi comparable à des spots où la profondeur est plus conséquente (sous-type épithéliale). Enfin, par souci de cohérence pour la visualisation, tous les scores CNA ont été ramené à une échelle entre -1 et 1.

## 2.5.2 Définition des altérations génomiques divergentes

**Tests statistiques** Enfin, afin de tester la significativité des altérations génomiques divergentes entre les différents sous-types tumoraux pour chaque patient, nous avons regardé si la distribution des scores CNA de toutes les cytobandes mineures pour chaque bras chromosomique (par exemple le bras p du chromosome 1 : 1p) était significativement différente de la distribution des scores CNA de toutes les autres cytobandes du génome pour chaque patient. Si les distributions étaient normalement distribuées (p-value d'un test de Shapiro supérieur à un seuil 0.05), alors un test t de Student était réalisé, et un test non paramétrique de Wilcoxon dans le cas contraire. Une correction de Bonferroni a été appliquée sur chacune des p-value du test statistique afin de corriger les biais liés aux test multiples. Les bras chromosomiques ont été considérés comme significativement différents entre compartiments si la p-value corrigée était inférieur à un seuil de 0.001 .

(Pour l'instant pas de seuil fload-change, a voir si je peux faire le VolvanoPlot avec la figure 10 et appliquer un seuil)

## 2.6 Software et packages

### 2.6.1 RStudio et language R

**Version du logiciel** L'ensemble des analyses bio-informatiques, ainsi que la génération des figures ont été réalisés à l'aide de scripts en R, développés dans l'IDE (Environnement de Développement Intégré) RStudio (version 2024.09.0+375 « Cranberry Hibiscus ») [19].

### 3 Résultats

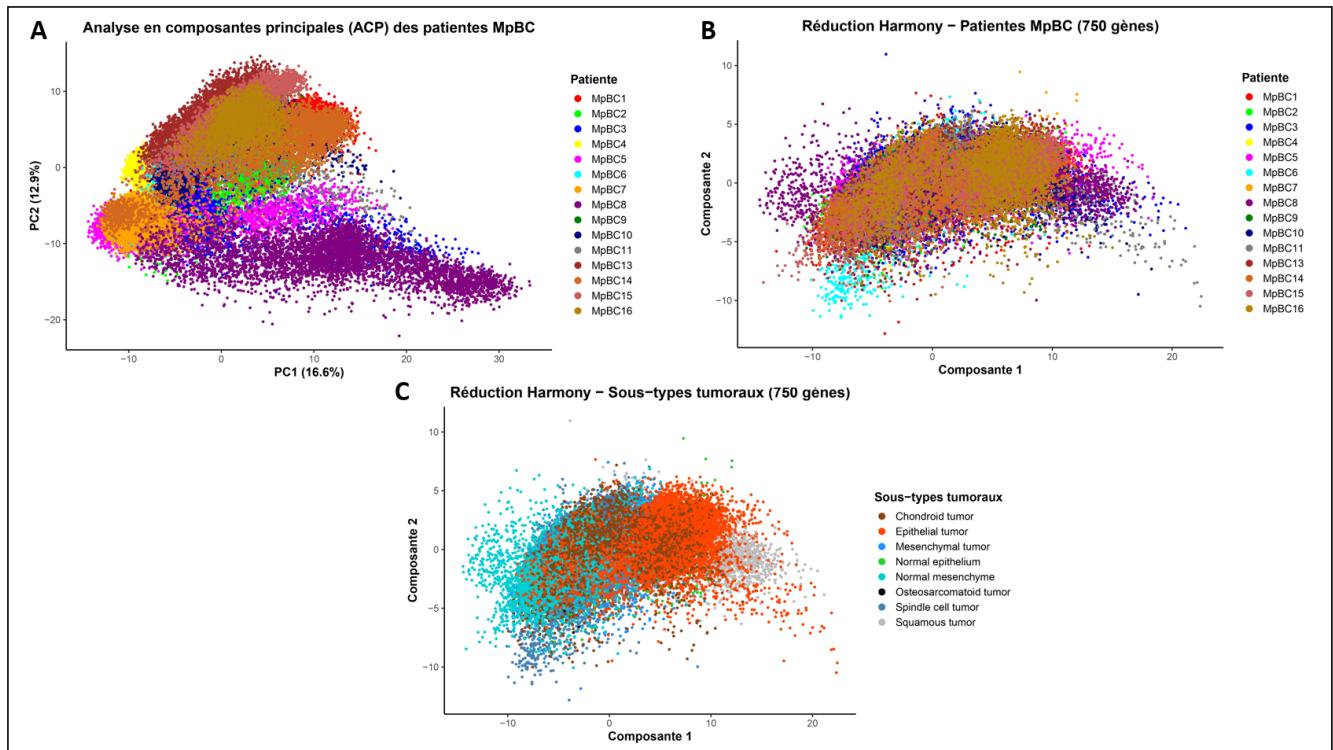
#### 3.1 Analyse des marqueurs phénotypiques

##### 3.1.1 Réduction de dimensionalité

Manipuler les matrices de comptes des données transcriptomiques Visium pour 16 patients est complexe. Afin d'analyser le profil transcriptomique de chaque échantillon, une première étape de réduction de dimension linéaire a été réalisée grâce à une ACP. La **figure 2A** représente les deux premiers axes cette réduction ACP. Chaque point représente un spot de cellules présent chez une des patientes MpBC, et ces points sont colorés selon la patiente d'origine. Ces 2 premières composantes permettent, ensemble, de résumer près de 30% de la variance totale présente dans le jeu de donnée. On voit également très bien que la position des spots dans l'espace des dimensions du jeu de donnée est dépendante de la patiente d'origine or cela peut biaiser les analyses en aval. Nous avons donc corrigé cet effet batch, dépendant du patient, avec la méthode Harmony. Dans notre analyse, nous avons également tenu compte d'autres effets confondants comme ceux liés au batch de séquençage et au numéro de la lame de séquençage utilisée. On constate avec la **figure 2B** que les spots colorés, ne se séparent plus selon les patients et se superposent en une forme similaire pour chaque patient. Cela suggère que la réduction de dimension corrigée par Harmony a bien fonctionné et permet donc d'effacer les différences entre les spots liés aux patients d'origine, qui pourraient biaiser notre analyse des véritables différences biologiques. La **figure 2C** nous montre cette réduction de dimension linéaire corrigé par Harmony selon tous les sous-types tumoraux retrouvés chez nos 16 patients. Bien que les biais techniques (batch de séquençage, lame utilisés, patient d'origine) aient été corrigé, on s'aperçoit avec cette **figure 2C**, que les différences biologiques dans notre jeu de donnée (représenté par les sous-types tumoraux) sont toujours capturées. En effet, on constate clairement sur la première composante de cette réduction, un axe épithélio-mésenchymateux correspondant au spectre de trans-différenciation retrouvé dans nos différents échantillons. Ainsi, avec cette étape de réduction de dimension suivie d'une correction par Harmony, nous avons réduit l'espace de dimension de notre jeu de donnée en 50 composantes, évitant les biais techniques tout en conservant les différences biologiques que l'on souhaite étudier.

##### 3.1.2 Clustering et enrichissement archétypal

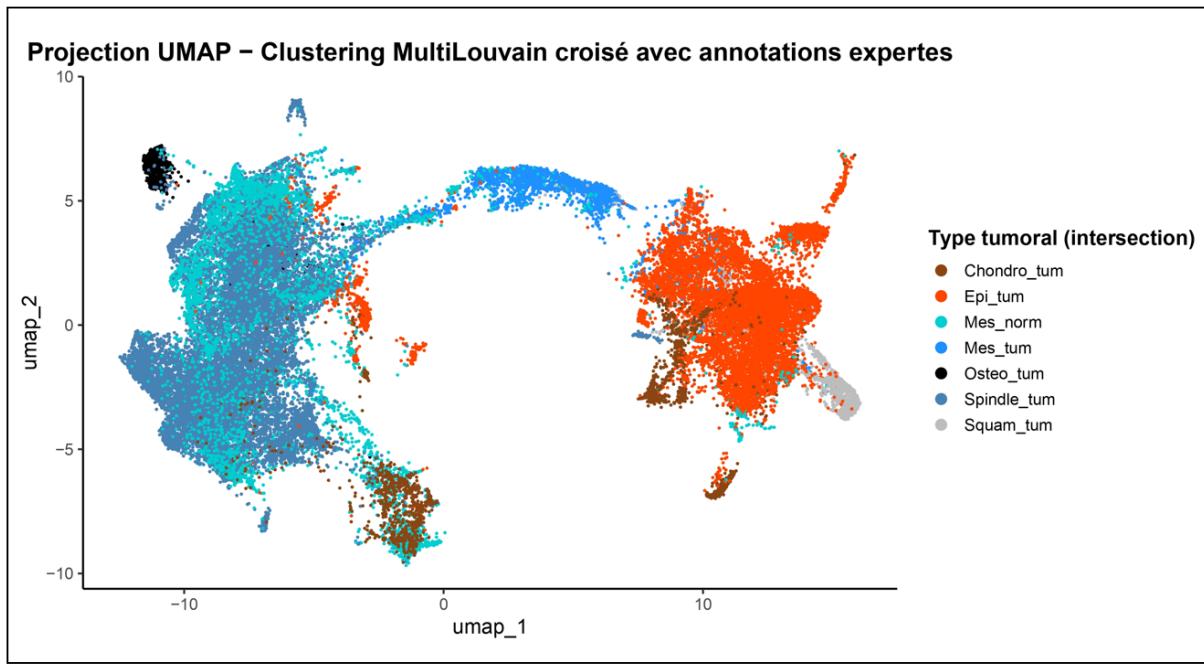
Nous avons utilisé les composantes principales corrigées des effets de lots issues de Harmony comme base de projection pour la visualisation UMAP. La UMAP est ici préféré à la t-SNE car elle préserve mieux les structures locales tout en maintenant la séparation des structures locales. Le clustering Multi-Louvain a permis de déterminer les clusters de spot de cellules dans l'espace vectoriel de la UMAP nous permettant de déterminer plusieurs clusters représentant les différentes annotations de sous-types tumoraux. Chacun de ces clusters ont été purifié des spots de cellules dont les annotations rentraient en contradiction avec l'annotation phénotypique de l'expert anatomopathologique. La combinaison de ces informations a permis d'obtenir la **figure 3**, représentant la projection UMAP des différents clusters de spot de cellule, correspondant à nos différents sous-types tumoraux retrouvés dans nos 16 échantillons MpBC. On constate à nouveau une première composante, discriminant les sous-types tumoraux selon un axe épithélio-



**Figure 2 : Correction du batch effect entre les 16 échantillons distincts de MpBC.** Représentation des deux premiers axes de l'analyse en composantes principales (ACP) des données transcriptomiques, avant correction (A) et après correction par Harmony, avec une coloration des points selon les patients (B) ou les sous-types tumoraux (C). Après correction, la première composante permet de distinguer les sous-types le long d'un axe épithélio-mésenchymateux.

FIGURE 2 – Correction du batch effect entre les 16 échantillons distincts de MpBC

mésenchymateux. On retrouve bien le cluster des cellules tumorales malpighiennes (Squam-tum) avec un profil « plus épithéliale » que les cellules tumorales épithéliales (Epi-tum). On peut également voir un cluster osteosarcomatoid bien distinct des autres clusters et un cluster de spot de cellules annotés fusiform (Spindle-tum) qui sont, sur l'axe, du côté opposé aux spot de cellules avec un profil épithéliales. Cela reflète plutôt bien la biologie observé des cellules tumorales dans les MpBC. Toutefois, on constate que certains clusters restent difficiles à caractériser et se fragmentent en plusieurs sous-clusters. C'est notamment le cas pour le cluster des cellules tumorales fusiformes (Spindle-tum) et tumorales chondroïdes (Chondro-tum), qui sont plus difficiles à caractériser. Enfin, les spots annotés associés aux cellules tumorales mésenchymateuses (Mes-tum) se positionnent entre les 2 gros regroupements de cellule considérés comme épithéliales et conjonctives, tandis que les cellules mésenchymateuses normales se répartissent et « contaminent » un peu tous les autres clusters. Cela est notamment dû aux limites de la technologie Visium, qui capture par spot plusieurs cellules et donc intègrent parfois des cellules de types cellulaires différentes de l'annotation attribué pour le spot. Pour conclure, cette projection UMAP faite à partir des données transcriptomiques des 16 patients MpBC permet de caractériser 6 clusters de cellules tumorales correspondant à nos 6 types de transdifférenciation dans les MpBC. Ces clusters pourront par la suite être utilisé pour déterminer des marqueurs phénotypiques (gènes) spécifique à chacun d'eux.



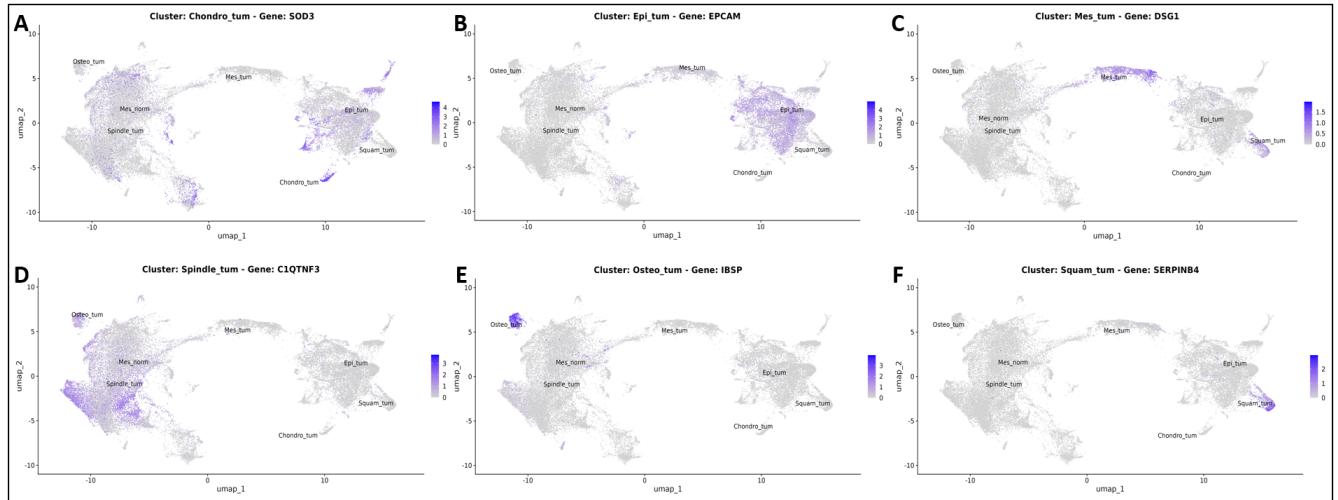
**Figure 3 :** Projection UMAP des données transcriptomiques Visium. La projection UMAP a été réalisée à partir des données transcriptomiques corrigées (Harmony), avec une coloration selon les sous-types tumoraux : **Chondro\_tum** = cellules tumorales chondroïdes, **Epi\_tum** = tumorales épithéliales, **Mes\_norm** = cellules mésenchymateuses normales, **Mes\_tum** = mésenchymateuses tumorales, **Osteo\_tum** = ostéosarcomatoïdes, **Spindle\_tum** = fusiformes, **Squam\_tum** = tumorales malpighiennes. Les sous-types cellulaires ont été obtenus par croisement des résultats d'un clustering MultiLouvain sur les données transcriptomiques et des annotations histopathologiques d'un expert. Les spots présentant des discordances entre ces deux sources (environ 15% des spots totaux) ont été exclus de la visualisation UMAP et des analyses ultérieures.

FIGURE 3 – Projection UMAP des données transcriptomiques Visium

### 3.1.3 Identification des gènes marqueurs

A partir des clusters de spot de cellule identifiés précédemment, et représentant les différents sous-types tumoraux caractéristiques des MpBC, nous avons effectué une analyse des gènes différentiellement exprimé entre ces clusters. En utilisant l'algorithme MAST, nous avons notamment identifié plusieurs marqueurs phénotypiques étant spécifiquement sur-exprimés dans chacun des clusters. Les plus significatifs et biologiquement intéressant d'entre eux sont illustrés via la projection UMAP de la **figure 4**. Tout d'abord, on constate que certains clusters sont plus difficiles que d'autre à caractériser via un gène qui soit spécifique de ce cluster. C'est le cas, par exemple, pour le cluster représentant les cellules tumorales chondroïdes (**figure 4A**), les cellules tumorales mésenchymateuses (**figure 4C**), et les cellules tumorales fusiformes (**figure 4D**). En effet, pour ces sous-types tumoraux, les marqueurs les plus spécifiquement sur-exprimé, respectivement SOD3, DSG1 et C1QTNF, se retrouve soit exprimé dans plusieurs autres clusters (marqueur non spécifique), soit spécifique uniquement d'une partie du cluster d'intérêt et pas la totalité du cluster (marqueur partiellement représentatif). Toutefois, on peut constater ce n'est pas le cas pour les autres sous-types tumoraux, où les marqueurs sont exprimés dans l'entièreté du cluster représentant le sous-type tumoral (représentatif de toutes les cellules du cluster) mais également avec une importante surexpression comparée aux autres clusters (surexpression spécifiques). De plus, la fonction de certains de ces marqueurs ont un lien avec le phénotype caractéristique de chacun de ces sous-types. Par exemple, la surexpression du marqueur EPCAM (Epithelial Cell

Adhesion Molecule) dans les cellules tumorales épithéliales (**figure 4B**) montrent que notre analyse arrive à capturer des marqueurs phénotypiques qui ont un réel sens biologique, en lien avec la biologie du sous-type tumoral concerné. Il en va de même pour le marqueur IBSP (glycoprotéine de la matrice extracellulaire osseuse) chez les cellules tumorales osteo-sarcomatoïdes (**figure 4E**), et le marqueur SERPINB4 (antigène connu des cancers kératinocytaires) pour les cellules tumorales malpighiennes (**figure 4F**). Pour conclure, on constate que certains clusters sont difficilement caractérisables par un marqueur phénotypique unique tandis que certains autres clusters, comme ceux représentant les sous-types tumoraux osteo-sarcomatoïdes, épithéliales et malpighiennes présentent des marqueurs surexprimés spécifiquement pour ces sous-types et sont pertinents sur le plan biologique.

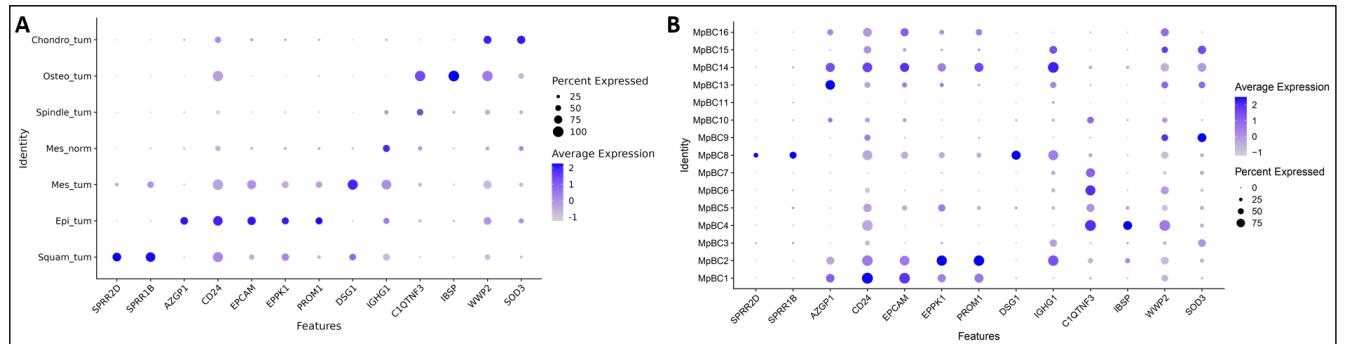


**Figure 4 :** Visualisation de l'expression génique par cluster sur la projection UMAP. La projection UMAP représente la distribution spatiale des clusters définis à partir des données transcriptomiques corrigées. Chaque panneau montre l'expression normalisée d'un gène d'intérêt au sein des cellules tumorales chondroïdes (A), tumorales épithéliales (B), tumorales mésenchymateuses (C), fusiformes (D), ostéosarcomatoïdes (E), tumorales malpighiennes (F), selon une échelle de couleur (du gris pour une faible expression au bleu pour une expression élevée). L'expression de ces gènes est enrichie de manière spécifique dans certains clusters, permettant l'identification de signatures transcriptomiques caractéristiques.

FIGURE 4 – Visualisation de l'expression génique par cluster sur la projection UMAP

Afin de comprendre plus en détail, l'expression de ces marqueurs dans ces clusters, et leur pertinence pour notre problématique, nous avons regardé l'expression moyenne et le pourcentage exprimé dans les cellules des meilleurs marqueurs, selon chacun des clusters identifiés (**figure 5A**) ou bien selon le patient d'origine (**figure 5B**). Nous pouvons constater sur la **figure 5A** que certains de ces marqueurs, comme le gène EPCAM, conforte l'idée qu'il s'agit d'un candidat intéressant pour caractériser le cluster des cellules tumorales épithéliales, car il est faiblement exprimé dans les autres clusters. De plus, on constate que ce marqueur est surexprimé pour la majorité des patients présentant ce type de transdifférenciation dans les tumeurs. De la même façon, d'autre marqueurs comme le gène SPRR1B, pourrait paraître être un marqueur pertinent pour caractériser le cluster des cellules tumorales malpighienne (**figure 5A**), cependant, en regardant la **figure 5B**, on constate que ce marqueur n'est exprimé uniquement pour 1 seul patient (MpBC8). Or le patient MPBC3 exprime aussi ce sous-type tumoral malpighien, et on ne retrouve pas d'expression de ce marqueur pour ce patient. Ainsi donc, cela nous montre donc que certains marqueurs identifiés, comme EPCAM, semblent être de bons candidats pour caractériser spécifiquement l'un des sous-types tumoraux, toutefois certains marqueurs de notre

analyse, retrouvé significativement surexprimé dans des clusters, semblent en réalité patient-dépendant.



**Figure 5 :** Profil d'expression des gènes marqueurs sélectionnés selon les types cellulaires tumoraux et les patients atteints de MpBC. Représentation DotPlot de l'expression normalisée de gènes sélectionnés, associés aux différents clusters des sous-types tumoraux (A) ou répartis parmi les 16 patients (B). Chaque point représente l'expression moyenne d'un gène dans un type cellulaire donné (A) ou chez un patient (B), la taille du point indiquant la proportion de cellules exprimant le gène, et la couleur du point reflétant l'intensité moyenne d'expression normalisée.

FIGURE 5 – Profil d'expression des gènes marqueurs sélectionnés selon les types cellulaires tumoraux et les patients atteints de MpBC

## 3.2 Analyse des altérations génotypiques divergentes

### 3.2.1 Profils génomiques biphasiques

**Réduction par cytobande** Pour obtenir le profil de CNV par patient représenté en **figure 6**, nous avons utilisé les résultats fournis par InferCNVplus. Les résultats fournis par ce package correspondent à des matrices représentant les variations du nombre de copies (CNV), où chaque ligne correspond à un gène et chaque colonne correspond à un spot. À chaque position est attribué un score d'altération du nombre de copie par rapport à un groupe de cellule de référence. Ce score est compris entre -1 et 1 : 0 signifie qu'il n'a pas été détecté de CNA pour ce gène dans les cellules tumorales, par rapport aux groupes de référence, +/- 1 signifie que ce gène est très probablement sujet à un CNA (délétion si score négatif, amplification si score positif). Pour obtenir le profil de CNV par patient représenté en **figure 6**, nous avons regroupé, pour chaque patient individuellement, les scores CNA des gènes des matrices InferCNVPlus appartenant aux mêmes cytobandes mineures, dans le génome humain. Afin d'être plus robuste et moins sensible face aux valeurs extrêmes, nous avons réalisé la médiane de ces scores CNA, plutôt que la moyenne. Cela permet notamment de réduire le nombre de données CNA en cytobandes mineures, tout en restant représentatif de la majorité des données. De plus, en lissant le signal CNV par région chromosomique, cela nous a permis d'éviter que certaines cellules ou gènes hautement amplifiés (ou délétés) ne faussent le signal.

**Normalisation par patient** Nous travaillons ici avec des spots de cellules, contrairement à ce qu'on a classiquement avec des cellules uniques. De plus il est important à annoter que chaque spot aura des profondeurs de séquençage différents, selon le sous-type tumoral considéré. En effet, nous devons donc procéder à une étape de normalisation afin de corriger les scores CNA des spots entre eux. Pour cela, à partir des médianes de scores CNA par cytobandes mineures pour chaque

patient, nous avons identifié le minimum (délétion maximale) et le maximum (amplification maximale) de chacun des 2 compartiments tumoral de l'échantillon bi-phasic. Nous avons calculé un facteur d'ajustement basé sur le rapport des extrêmes entre les 2 tissus et ajusté les scores CNA du tissu « le moins extrêmes ». Par exemple, pour le patient MpBC9 de la figure 5, c'est le tissu tumoral épithelial qui présentait la plus forte amplification (**figure 6A**, chromosome 8), on a donc ajusté les amplifications (scores CNA > 0) du tissu appairé, le tissu tumoral chondroïde. La même transformation a été appliquée pour les délétions (scores < 0). Enfin, afin de faciliter la visualisation, les scores ont été ramenés sur une échelle de -1 à 1. Cette normalisation a donc permis, visuellement de faire coïncider les extreums des scores de délétion et d'amplification entre les différents compartiments tumoraux (**figure 6B**, chromosome 8), et, in fine, de limiter les biais aux différences de profondeurs de séquençage entre les différents spots annoté pour les différents sous-types tumoraux.

### 3.2.2 Identification des altérations divergentes entre compartiments pairés

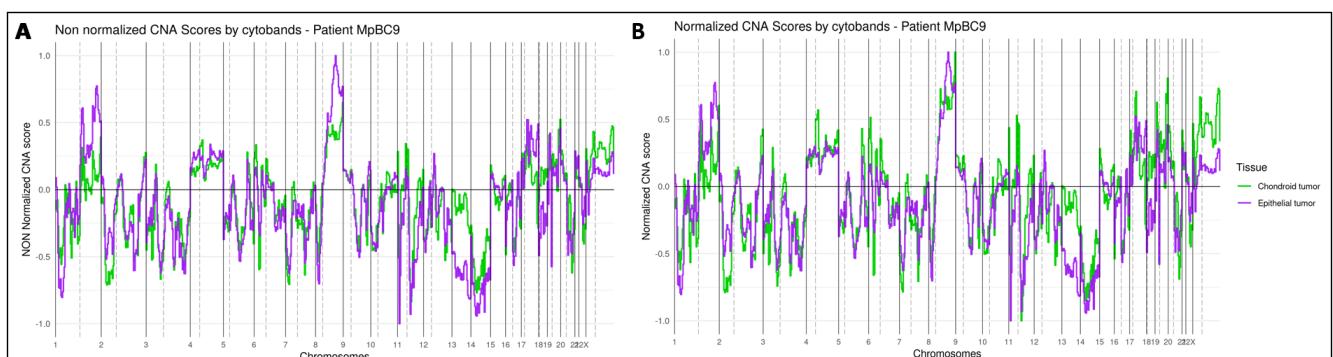
**Test par bras chromosomique** Afin d'identifier les altérations divergentes entre les compartiments pairés au sein de chaque tumeur chez nos 16 patientes, nous avons testé si la distribution des scores CNA pour un bras chromosomique donnée, était significativement différente des scores CNA des autres cytobandes, pour chaque patient individuellement. Pour cela nous avons défini pour chaque cytobande mineure, le bras chromosomique qui lui est associés (bras chromosomique p ou bras chromosomique q). Ainsi, grâce un test t de Student nous avons pu comparer pour chaque patient, la distribution des cytobandes de chacun des bras chromosomiques et tester si l'un de ces bras étaient significativement différents de la distribution de autres cytobandes. NOus avons ainsi obtenue une p-value pour chaque bras chromosomique.

**Correction multiple** En réalisant les tests statistiques successifs, il est nécessaire de réaliser des corrections multiples pour contrôler le taux de faux positif, dont la probabilité augmente avec le nombre de test réalisé. Dans le cas de notre problématique nous avons appliqué une méthode de correction stricte, correspondant à la correction de Bonferroni. Ainsi les p-valeures obtenues lors des tests ont été corrigé afin de fournir des p-valeures ajustées qui permettent une interprétation non biaisée des résultats, avec une probabilité de commettre une erreur extrêmement faible, ce qui souhaitable dans le cas de recherche clinique comme dans notre projet.

**Résultats significatifs** Pour chaque patient, afin de définir si le profil CNA d'un bras chromosomique est significativement différent du profil des autres cytobandes, nous avons utilisé un seuil de 0.001. En plus de ce seuil alpha, une 2e métrique à été utilisé correspondant à la taille de l'effet. Cette taille de l'effet a été réalisé pour chaque bras chromosique de chaque patient. Elle consiste en la différence des médianes des scores CNA des cytobandes par bras chromosomiques, entre les tissus pairés. Cela permet ainsi de mesurer l'écart des scores médians entre les 2 tissus, pour un bras donné et un patient donné. En effet, en utilisant cette métrique comme filtre, nous pourront éliminer les bras chromosomiques significatifs mais ayant une taille d'effet trop peu importante, et donc potentiellement moins intéressants.

### 3.2.3 Visualisation des différences de score CNA entre compartiment

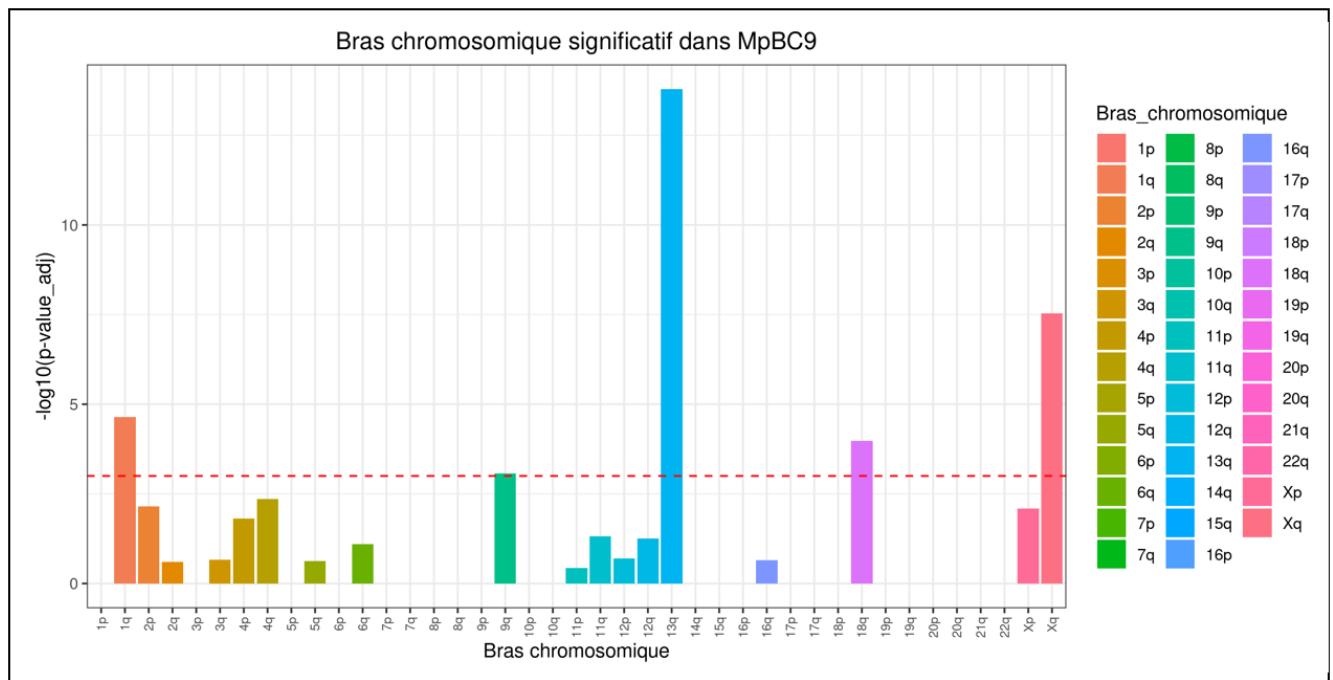
La **figure 6**, permet de représenter les scores CNA obtenues par InferCNVPlus, de chacune des cytobandes mineures pour le patient MpBC9. Sur les 2 panels de la figure, la couleur des courbes correspond aux compartiments tumoraux analysés chez ce patient. On peut tout d'abord voir que l'amplitude des scores CNA pour les scores non normalisés du tissu tumoral épithéial est moins grande que pour le tissu pairé, le tissu tumoral chondroïde (**figure 6A**). Dans la majorité des cytobandes, c'est à chaque fois le compartiment épithéial qui est le plus élevé comparé à l'autre compartiment. Ces différences sont typiques d'une profondeur de séquençage différente entre les 2 compartiment. La normalisation (**figure 6B**) va permettre d'éliminer ce biais et d'analyser les réelles différences biologiques existant entre les 2 compartiments tumoraux. On peut voir que l'ajustement des scores efface certaines différences entre les 2 tissus, qui étaient présente sans normalisation. Par exemple, il semblait il y avoir une importante différence de score CNA pour le bras chromosomique 8q avant normalisation, et qui a été complètement corrigé avec la normalisation. En revanche, on peut remarquer que d'autres différences présentes avant la normalisation, tel que le bras 13q, sont toujours présente même après normalisation. Cela nous informe que les différences de score CNA observées sur ce bras ne sont pas dû à une différence de profondeur, et qu'il y a probablement une autre raison expliquant cette différence entre les tissus tumoraux. Cette région chromosomique présente donc un intérêt certain et mérite des analyses plus approfondies.



**Figure 6 :** Représentation des scores CNA obtenus avec InferCNVPlus, avant (A) et après normalisation (B), en fonction des cytobandes mineures pour chaque chromosome du génome. La ligne verticale en pointillé indique la position du centromère pour chaque chromosome. Les couleurs des courbes correspondent aux deux sous-types tumoraux comparés chez le patient MpBC9 (Chondroid tumor et Epithelial tumor). Les scores CNA s'étendent de -1 à 0 (délétion) et de 0 à 1 (amplification). La normalisation permet de corriger les biais liés au nombre de cellules par spot et liés à la dilution du signal CNA dans les spots de cellules des différents sous-types tumorales.

FIGURE 6 – Représentation des scores CNA obtenus avec InferCNVPlus, avant (A) et après normalisation (B), en fonction des cytobandes mineures pour chaque chromosome du génome

Les tests statistiques réalisé par bras chromosomiques permettent d'identifier les régions chromosomiques significativement altéré comparé aux autres régions du génome chez le patient. On peut ainsi, à partir de la **figure 7**, visualiser les régions génomiques intéressantes et à investiguer. Dans cette figure, nous pouvons voir que les bras chromosomiques présentant des scores CNA significativement différents du reste des régions génomique sont les bras 1p, 13q et Xq. En effet, la p-valeurs associées à ces bras chromosomiques sont supérieur à notre seuil alpha de 1% fixée.

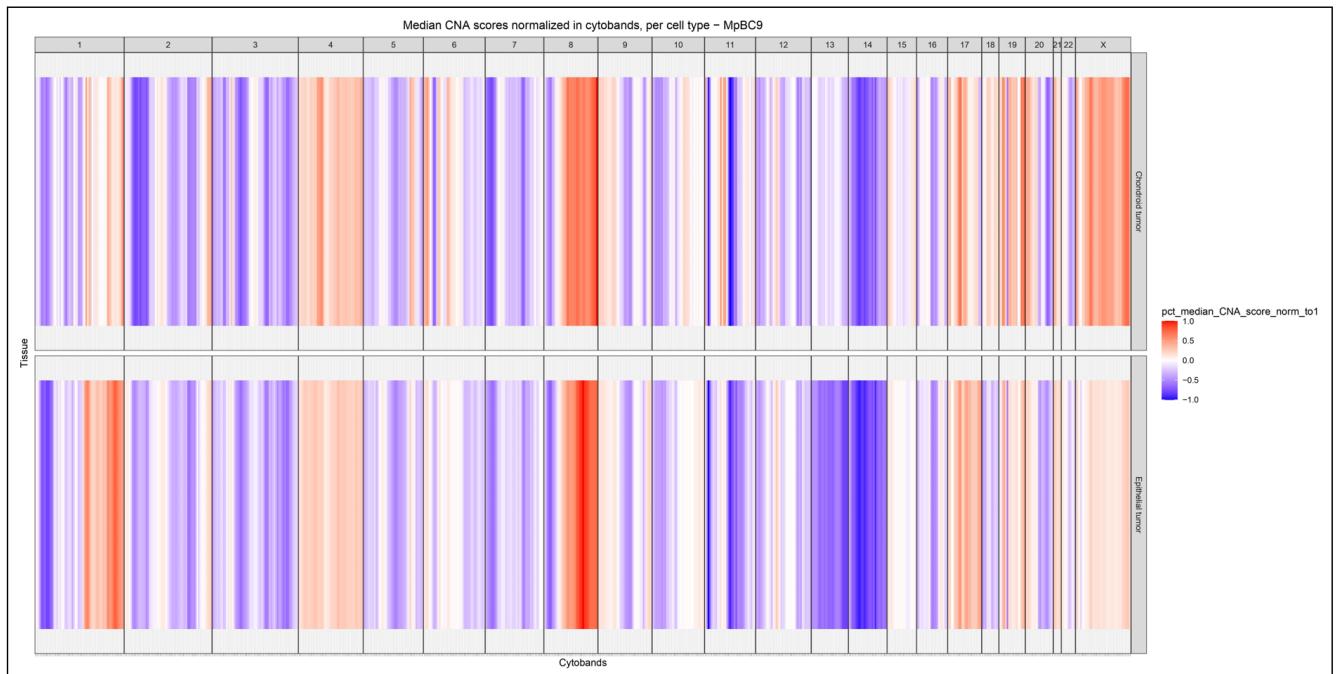


**Figure 7 : Profil de significativité des altérations chromosomiques pour le patient MpBC9.** Représentation des valeurs  $-\log_{10}(p\text{-value\_ajustée})$  pour chaque bras chromosomique. La ligne rouge horizontale en pointillés indique le seuil de significativité fixé à  $\alpha = 0,001$  (0,1 %). Les bras 1q, 13q et Xq présentent les niveaux de significativité les plus élevés chez le patient MpBC9 et constituent des régions d'intérêt potentiel.

FIGURE 7 – Profil de significativité des altérations chromosomiques pour le patient MpBC9

Une manière plus intuitive de voir ces résultats est avec une carte de chaleur, aussi appelé *heatmap*, comme sur la **figure 8**. Ici, on peut observer pour chaque chromosome du génome, les scores CNA associées à chaque cytobandes et coloré selon un spectre de couleur avec les amplifications en rouge et les délétions en bleu. On retrouve notamment, l'importante amplification en 1q des cellules tumorales chondroïdes par rapport aux cellules tumorales épithéliales, mais également la très importante délétion en 13q pour les cellules tumorales épithéliales et l'importante amplification en Xq pour les cellules tumorales chondroïdes.

Enfin sur la **figure 9**, nous avons tracé le graphique montrant la corrélation des scores CNA des cytobandes pour les 2 compartiments tumoraux retrouvés dans l'échantillons MpBC9. Chaque point représente le score normalisé obtenu pour chaque cytobande. Les différents panels de la figure mettent en évidence les cytobandes appartenant au bras retrouvé comme significatif par le test statistique (point coloré en rouge). On constate que pour les 3 bras chromosomiques significatif chez MpBC9 (1q en **figure 9A**, 13q en **figure 9B** et Xq en **figure 9C**), les cytobandes appartenant à ces bras sont outliers par rapport aux autres points et dévient de la droite de corrélation parfaite (droite noire en trait plein). Cela soutient une nouvelle fois que ces régions génomiques sont dignes d'intérêt, en tout cas pour ce patient MpBC9. On constate également que certaines cytobandes des bras chromosomiques significatifs (points en rouges) ne sont parfois pas toutes outliers. Cela implique donc peut-être des altérations du nombre de copie à

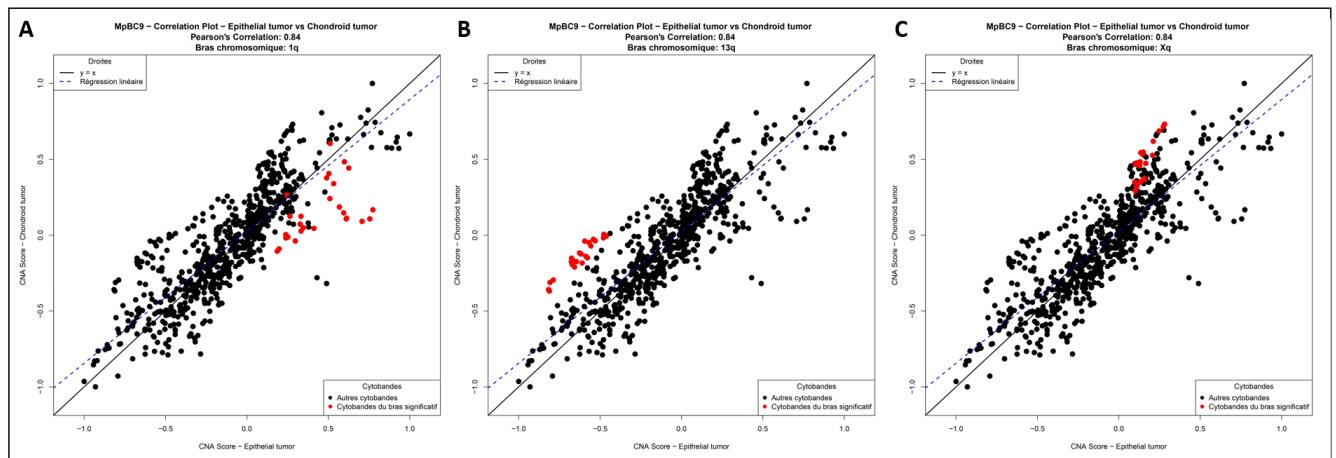


**Figure 8 :** Carte de chaleur des altérations du nombre de copies pour le patient MpBC9. Heatmap représentant les scores CNA normalisés pour chaque sous-type tumoral comparé chez le patient MpBC9 (chondroid tumor vs epithelial tumor), selon les cytobandes mineures réparties le long des chromosomes 1 à 22 et X. L'intensité de la couleur reflète le degré d'altération du nombre de copies, après normalisation. On retrouve notamment des altérations significatives dans les régions chromosomiques d'intérêt 1q, 13q et Xq.

FIGURE 8 – Carte de chaleur des altérations du nombre de copies pour le patient MpBC9

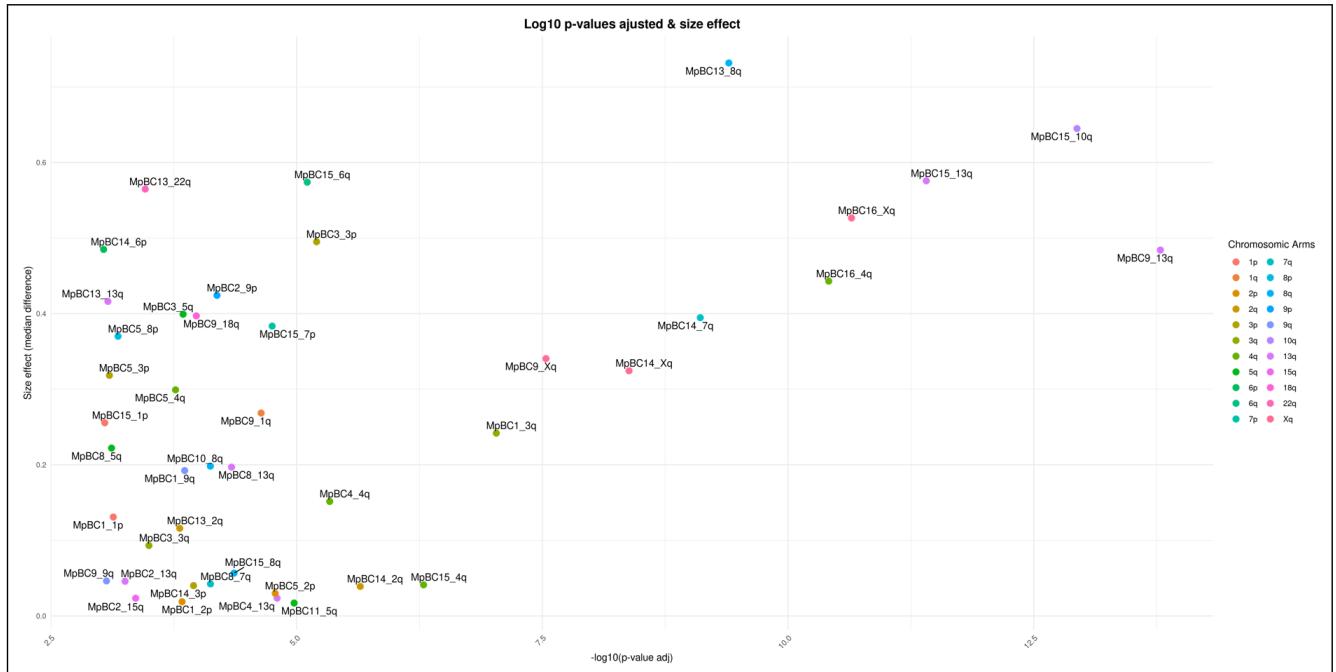
une échelle plus petite que les cytobandes mineures du génome, et des altérations plus localisées.

Ce travail a été réalisé pour chacun des 16 patients, et les tests statistiques ont été appliqués par bras chromosomiques en calculant à chaque fois la taille de l'effet, c'est-à-dire la différence médiane des scores des cytobandes par bras chromosomique. La **figure 10** récapitule tous ces bras chromosomiques significatifs selon cette taille d'effet pour l'entièreté de notre dataset (toutes les patientes). Parmi les premiers bras chromosomiques les plus significatifs et avec une taille de l'effet importante, on peut remarquer la présence récurrente des bras 13q et du bras Xq, on retrouve ces 2 bras chromosomiques comme très largement significatif comparé aux autres bras, au seuil de 0.1%. Par exemple sur les 4 patients dans lequel nous comparons le compartiment tumoral épithéial contre tumoral chondroïde, 3 d'entre eux, présente le bras chromosomique 13q comme disposant d'une altération du nombre de copie significatif. Cela nous donne des pistes de recherche afin de poursuivre l'analyse au niveau de ces régions chromosomiques et comprendre les possibles causes génomiques des MpBC.



**Figure 9 : Analyse de la corrélation des scores CNA entre les deux sous-types tumoraux du patient MpBC9.** Représentation de la corrélation entre les scores CNA d'un tissu chondroid tumor (ordonnée) et du tissu epithelial tumor (abscisse) pour le patient MpBC9. Chaque point correspond à une cytobande mineure d'un bras chromosomique. Les points colorés en rouge indiquent les cytobandes situées sur les bras chromosomiques identifiés comme significativement altérés précédemment : 1q (A), 13q (B) et Xq (C). La ligne noire pleine représente une corrélation parfaite (droite d'identité), tandis que la ligne pointillée bleue indique la droite de régression linéaire ajustée au nuage de points. Les cytobandes des bras significatifs apparaissent comme des outliers, s'éloignant nettement de la droite d'identité, ce qui suggère une véritable altération différentielle du nombre de copies dans ces régions.

FIGURE 9 – Analyse de la corrélation des scores CNA entre les deux sous-types tumoraux du patient MpBC9



**Figure 10 : Taille de l'effet des altérations chromosomiques entre sous-types tumoraux.** Pour chaque patient, la médiane des différences de scores CNA entre les deux tissus comparés est représentée en ordonnée pour chaque bras chromosomique. L'axe des abscisses indique la significativité statistique associée à ces différences, exprimée en  $-\log_{10}(p\text{-value adj})$ . Chaque point correspond à un bras chromosomique pour un patient donné.

FIGURE 10 – Taille de l'effet des altérations chromosomiques entre sous-types tumoraux

## 4 Discussion

### 4.1 Limites de la technologie Visium (sous-titres à supprimer)

#### 4.1.1 Risque d'hétérogénéité cellulaire intra-spot

**Séparation imparfaite des clusters (UMAP)** La technologie de transcriptomique spatiale Visium proposé par 10X Genomics, repose sur un séquençage d'un échantillon en préservant l'information de la provenance spatiale de chaque reads. Chaque lecture est attribuée à un spot spécifique correspondant à une région précise de l'échantillon. Ces spots ont une taille d'environ 55µm et peuvent contenir jusqu'à 10 cellules simultanément par spot. Bien que cette technologie permette d'allier l'information du séquençage transcriptomique à la structure spatiale du tissu, ce qui n'est pas le cas en faisant du bulk ou du single cell, cette dichotomie du tissu par spot comporte certaines limitations. Il est vrai, une première limite de cette technologie est que les ARN peuvent diffuser dans le tissu avant d'être captés et séquencé, et donc amené à une contamination des spots entre eux. Par exemple le signal tumoral de certains spots peut « déborder » sur des zones non tumorales proches. Dans notre projet, où un échantillon contient plusieurs sous-types tumoraux, cela amène donc à des contaminations locales très difficiles à éliminer et un bruit de fond compliquant les analyses des profils transcriptomiques en aval. De plus, le principe de séquençage par spot de cellule comme le fait Visium, amène à une résolution assez limitée, où chaque spot va capturer les ARN de plusieurs cellules en même temps. Avec les échantillons MpBC biphasiques tel que nous avons (qui contiennent au moins 2 compartiments tumoraux ou plus), nous avons utilisé les connaissances d'une experte en anatomopathologie pour annoter au mieux le type tumoral de chaque spot selon l'histologie, la morphologie, le phénotype des cellules contenue dans chacun d'eux. L'objectif étant d'utiliser dans notre analyse des spots de cellules où nous sommes certains du sous-type tumoral des cellules capturées, et d'obtenir un marqueurs phénotypique spécifique de chaque sous-type. Cependant, il n'est pas possible d'être certains que nos spots annotés pour un sous-type tumoral ne contiennent pas quelques cellules non tumorales (cellules stromales normales, immunitaires), voire des cellules tumorales d'un autre sous-type. En conséquence, ces imprécisions rend difficile l'attribution précise d'un signal à un sous-type tumoral spécifique, et se traduit donc un positionnement parfois diffus des spots dans l'espace vectoriel de la PCA et les projections UMAP. C'est par exemple le cas, pour le sous type fusiforme (spindle-like). Ces cellules tumorales ressemblent phénotypiquement à des cellules du tissu conjonctif et produisant de la matrice extracellulaire. Les spots de cellules de ce sous type sont difficiles à annoter et cela se traduit par une séparation assez floue entre les différents sous-types mésenchymateux sur la représentation de la réduction de dimension PCA et la projection UMAP. Il est difficile de définir une frontière exacte entre les différents sous-types mésenchymateux et les différents clusters se superposent entre eux. Cela n'est rien d'autre que le reflet de la contamination des spots annotés pour un sous type tumoral spécifique, par des cellules d'un autre sous-type tumoral.

La solution permettant de résoudre cette limitation de la technologie Visium et qui permettrait de connaître la composition cellulaire d'un spot et d'attribuer à chaque sous-type tumoral un profil transcriptomique précis est d'utiliser une technique de déconvolution. Le principe de la déconvolution est de considérer les spots de cellules comme une combinaison linéaire de profils

d'expression cellulaire. Ainsi, grâce à cette étape de déconvolution, nous pourrions connaître les sous-types tumoraux présent dans chaque spot et estimer leurs proportions à partir du profil transcriptomique globale du spot. Des travaux précédents au sein de l'équipe ont déjà montré que des outils, tel que RCTD (Robust Cell Type Decomposition), sont des algorithmes puissants pour la déconvolution de spot Visium et semble plutôt bien adaptés à la biologie des MpBC. Cependant, l'utilisation d'une déconvolution nécessite de comparer le profil transcriptomique mixte de chaque spot aux profils transcriptomiques de référence caractéristique de chaque type cellulaire. Or actuellement, aucun travail de recherche n'a encore construit de tel profil transcriptomique spécifique des différents sous-types tumoraux des MpBC. Je détaillerai plus en détail dans les paragraphes suivants, la manière dont nous construirons ces profils transcriptomiques de référence pour chaque sous-type.

**Trop faible spécificité des marqueurs phénotypiques** Actuellement, les contraintes qu'imposent la technique Visium sans déconvolution par spot, rend l'identification des marqueurs phénotypiques spécifiques de chaque sous-type compliqué. En effet, on constate que pour le moment, les marqueurs identifiés avec l'étape de DGEA identifiés pour chaque sous-type tumoral correspondent à des gènes étant encore trop patient-spécifique plutôt que cluster-spécifique. L'hypothèse expliquant cette difficulté vient probablement de la faible capacité résolutive des spots de cellule de la technologie Visium et de l'hétérogénéité cellulaire capturée par chaque spot. Cela rend notamment compliqué et imprécis le clustering des cellules tumorales selon leur profil transcriptomique. Nous nous retrouvons donc avec des marqueurs phénotypiques parfois surexprimé dans plusieurs clusters, ce qui n'en font pour l'instant pas des candidats satisfaisants, par rapport à notre problématique pour le diagnostic clinique des différents sous-type tumoraux. Toutefois, certains marqueurs phénotypiques identifiés montrent que nous arrivons en partie avec notre analyse à capturer une partie de la biologie représentant les différents sous-types tumoraux. Par exemple les marqueurs EPCAM et CD24 identifiés dans le cluster correspondant aux cellules tumorales épithéliales, codent pour des protéines membranaires impliquées dans les interactions cellules-cellules et cellules-matrice. De nombreuses études montrent aujourd'hui leur surexpression ainsi que leur implication dans la progression tumorale au sein de divers carcinomes épithéliaux (ref.).

Un deuxième axe d'amélioration afin d'identifier des marqueurs phénotypiques spécifiques des types tumoraux et moins dépendant de l'identité du patient, concerne l'étape de correction du batch-effect lors de l'intégration des données transcriptomiques de tous les patients. Il est vrai, une correction pas assez importante des comptes d'UMI dans les matrices de comptage de gène des patients peut aboutir à la production de marqueurs plus spécifiques aux patients qu'aux types tumoraux considérés. Le processus d'intégration de multiples de données et de multiples patients est une étape primordiale dans les analyses single-cell, elle n'en demeure toutefois pas moins compliquée. En effet, cette étape délicate nécessite de trouver le bon compromis dans les paramètres entre éliminer les différences propres aux patients sans toutefois sur-corriger les données et éliminer les réelles différences biologiques, au risque de perdre le signal d'intérêt. Dans notre analyse, l'identification de marqueurs spécifiques aux patients (**figure 5B**) suggère que l'intégration des 16 patients et la correction du batch-effect avec Harmony n'est pas pleinement satisfaisante et pourrait probablement être amélioré. Ainsi, des ajustements au niveau de

la méthodologie ou l'exploration d'outils alternatifs d'intégration pourrait permettre d'affiner encore la robustesse des marqueurs phénotypiques identifiées pour chaque sous-type tumoral.

**Dilution/Perturbation du signal CNA** Les limites imposées par la technologie Visium, notamment l'incorporation de cellules non tumorales au sein des spots de cellules annotés comme tumoraux, peut biaiser l'inférence du score CNA prédit par InferCNVPlus. L'estimation de ces scores CNA repose sur la détection d'expression de gène caractéristiques des altérations du nombre de copies génomiques. Or, l'incorporation de cellules non tumorales va « diluer » le signal tumoral, car les cellules non tumorales ont tendance à avoir un profil transcriptomique plus stable et bien différents des cellules tumorales. Par ailleurs, dans notre projet, tous les spots n'auront pas la même densité cellulaire ce qui dépendra du sous-type tumoral considéré. Par exemple, un spot contenant des cellules tumorales chondroïdes, isolées au sein de grosses logettes de matrice cartilagineuse, présentera une densité cellulaire plutôt faible. A l'inverse, un spot capturant des cellules épithéliales, de nature jointives et adhérentes entre elles, sera plus densément peuplé. En conséquence, le score d'altération du nombre de copies s'en retrouvera sous-estimé pour certains spots, faussant l'analyse. Cette hétérogénéité cellulaire intra-spot, intrinsèques aux spots Visium, doit donc être corrigé comme nous l'avons fait avec la normalisation selon les compartiments cellulaires au sein de chaque échantillon. Au final, cela a permis de ramener les scores CNA entre les compartiments d'un même échantillon à des niveaux comparables.

**Difficulté pour caractériser des marqueurs génétiques et CNA spécifiques aux sous-types** L'identification des marqueurs génétiques et les CNA spécifiques de chaque sous-type tumoral sont primordial pour comprendre les causes de l'hétérogénéité cellulaires au sein des MpBC. Grâce aux test statistiques, nous avons pu déterminer pour chaque patient des régions du génome représenté par des bras chromosomiques, significativement altéré par rapport au reste des autres régions du génome. Cependant, il est important à garder en tête que la méthode de détection des altérations génomiques proposée par InferCNVPlus repose sur le profil d'expression transcriptomique. C'est donc une approche indirecte, qui estime un profil génomique à partir des données transcriptomiques, et qui est sensible, au bruit, à la pureté et la composition cellulaires des « spots » annotés tumoraux. D'autre part, certains sous-types tumoraux sont représentés par un faible nombre de spot et parfois présent dans peu de patient. C'est par exemple le cas avec le sous-type ostéosarcomatoïde, qui une forme très rare de transdifférenciation dans les MpBC, et pour lequel nous n'avons actuellement dans notre dataset qu'un échantillon tumoral issus d'un seul patient (MpBC4). Cela amène à une puissance réduite lors du test statistique et donc à une probabilité moindre de détecter une altération du nombre de copie pour ce sous-type. Il serait donc intéressant pour la suite de refaire cette analyse génomique des altérations du nombre de copie avec les données transcriptomiques déconvolus par spot selon la composition cellulaires et en intégrant plus de patients de la cohorte, en priorisant ceux présentant des sous-types rares comme le sous-type ostéosarcomatoïde, afin d'avoir des résultats plus robustes et des marqueurs génétiques fiables.

**Les CNAs divergents identifiés sont-ils passagers ou drivers ?** Parmi tous les potentiels CNA divergents identifiés entre les différents compartiments pour chaque sous-type tumoral, la

véritable question est de savoir si ces altérations génomiques sont passagères ou au contraire drivers de la transdifférenciation tumorale chez les MpBC. Une mutation passagère correspondrait à une altération du nombre de copie sans réel rôle fonctionnel, apparaissant de façon aléatoire et ponctuel pour un patient, et qui disparaîtrait de la population tumorale au bout de quelques générations. En revanche, un CNA drivers correspond à un changement du nombre de copies de gène spécifiques contribuant activement à la tumorigénèse. Bien souvent, ces gènes tels que des oncogènes régulant la prolifération cellulaire présentent une amplification dans leur régions promotrices, tandis que les gènes suppresseurs de tumeurs vont présenter des délétions importantes, impactant leur activation et leur rôle dans la répression de la tumorigénèse. De plus, on peut légitimement admettre qu'un CNA drivers apparaît dans différents patients voire dans différents sous-types tumoraux des MpBC. En effet, si cette altération génomique est la cause de la transdifférenciation des MpBC, alors nous devrions retrouver ce CNA dans toutes les populations cellulaires tumorales qui découlent de cette transdifférenciation. Dans notre analyse, nous retrouvons par exemple le bras chromosomal 13q fortement délété pour les cellules tumorales épithéliales comparé aux cellules tumorales chondroïdes. Sur les 4 patients avec ce type de comparaison entre compartiment, on dénombre 3 d'entre eux qui présente cette délétion significative au niveau de cette région chromosomique 13q. La récurrence de cette altération du nombre de copie peut suggérer la présence d'une mutation driver dans cette région génomique et un potentiel rôle dans la transdifférenciation des cellules tumorales épithéliales. Dans les prochaines analyses, il serait particulièrement intéressant d'explorer les gènes présents dans cette région et d'évaluer les gènes qui pourraient constituer des drivers spécifiques de ce sous-type tumoral pour les MpBC. Cependant, pour des résultats réellement fiables et robustes, cela nécessite d'être confirmé dans une cohorte plus importante et incluant plus de patientes.

## 4.2 Perspectives du projet (sous-titres à supprimer)

### 4.2.1 Consortium MAESTRO : échantillons additionnels

La problématique de ce travail s'inscrit dans un cadre plus large d'un programme de recherche nationale sur le cancer du sein, avec la collaboration du CLB à Lyon, de l'institut Bergonié à Bordeaux et de l'Institut Curie à Paris. Le projet MAESTRO (MetaplAstic brEaST caRcinOma) vise spécifiquement à comprendre davantage les cancers du sein métaplastiques, rares et résistants aux traitements classiques. Ce consortium regroupe de nombreux chercheurs et praticiens cliniques, et centralise la collecte des données cliniques et génétiques des patientes atteintes de carcinome mammaires métaplastique en France. A ce jour, on dénombre 140 tumeurs de MpBC réunies grâce à ce projet, fournissant aux chercheurs un pool de données importantes pour développer la recherche de nouvelles thérapies et marqueurs de diagnostic afin d'améliorer la prise en charge des patientes atteintes de ces cancers métaplastiques [20]. La constitution de cette cohorte permet également de faciliter la création d'essais cliniques qui, dans le cadre des cancers rares comme les MpBC, est parfois difficile à réaliser. Enfin, dans le cadre de ce travail de recherche, le consortium MAESTRO nous donne l'opportunité d'enrichir notre dataset avec de nouveaux échantillons. Cela nous sera particulièrement utile et indispensable, notamment pour d'augmenter les effectifs et la puissance des analyses lors de nos tests statistiques. En effet nous avons vu précédemment lors de l'analyse des marqueurs phénotypiques et génétiques, que

certains sous-types tumoraux, comme les cellules ostéosarcomatoïdes, ne sont représentés que par une seule patiente dans notre dataset actuellement. Afin d'avoir plus de robustesse dans les marqueurs identifiés, il est nécessaire d'inclure des échantillons supplémentaires et les plus divers possibles.

#### 4.2.2 Analyse de réseaux / Pathways dérugglés plutôt que des gènes spécifiques

La problématique de notre travail est de trouver des marqueurs d'expression caractéristiques et spécifiques de chacun des sous-types tumoraux dans les MpBC. Nous n'avons donc, aujourd'hui pas ou peu, de piste sur lesquels nous reposer pour identifier des marqueurs fiables. De fait, devant la grande diversité et l'important nombre des gènes potentiels identifiés durant l'analyse DGEA des marqueurs d'expression de chaque sous-type tumoral, il est complexe de savoir quels marqueurs sont pertinents. Dans notre étude, afin de réduire le nombre de gène d'intérêt tout en gardant les gènes pertinents qui pourrait être faiblement exprimés, l'utilisation de logiciel comme SCENIC (Single-Cell rEgulatory Network Inference and Clustering) serait pertinente. En effet, cet outil nous permettrait de rechercher les régulateurs clés de plusieurs voies de signallisations d'intérêt chez les MpBC, et d'expliquer une expression différentielle de ces marqueurs dans chaque sous-type MpBC. Ainsi nous pourrions expliquer les profils d'expressions et trans-différenciation observés dans chaque sous-type, selon les facteurs de transcription exprimés et les potentiels voie de signalisation dérégulées. Enfin, pour garantir que ces marqueurs d'expression identifiés soient bien représentatifs de la majorité de patientes partageant les même sous type-tumoral, plutôt que spécifique pour quelques patients comme dans notre analyse, nous réaliserons les analyses en regroupant les échantillons partageant les mêmes histologies de trans-différenciation.

#### 4.2.3 Analyse approfondie via snRNA-seq

**Création d'un atlas MpBC spécifique** Nous avons pu voir que lors des analyses des marqueurs phénotypiques et génomiques, nous avons rapidement été confrontés aux limitations de la technologies Visium. En effet, le manque de résolution intrinsèquement liée aux spots utilisés par cette technique, amène à la présence d'hétérogénéité cellulaire au sein de certains spots, et donc à des difficultés caractériser le profil transcriptomique et génomique des différents sous-types tumoraux. Une solution, évoquée précédemment, serait de compléter l'analyse par une étape de déconvolution du signal transcriptomique pour chaque spot. Cela permettrait d'identifier ainsi les différents types cellulaires et leurs proportions au sein des spots. Cependant, pour réaliser cela, il est nécessaire d'avoir à disposition, un signal/profil transcriptomique de référence pour chacun des sous-types tumoraux des MpBC. Or, actuellement, il n'existe aucune référence pour ces types cellulaires spécifiques aux carcinomes métaplastiques du sein. Bien sûr, il existe de plusieurs ensembles et bases de données transcriptomiques capturant l'hétérogénéité inter et intra-tumoral des cellules cancéreuses dans les cancers du sein. Cependant, de précédentes analyses transcriptomiques de l'équipe, utilisant ces données pour faire de la déconvolution des spots Visium des échantillons MpBC, ont montré des résultats peu satisfaisants et non spécifiques. En effet, par exemple, le signal transcriptomique des spots de cellules tumorales annotés « fusiformes » (Spindle), étaient majoritairement attribué à la présence de fibroblaste. Les cellules

tumorales fusiformes des MpBC ont, certes, peut-être un profil d'expression transcriptomique se rapprochant des fibroblastes, mais ce sont bien deux entité cellulaire distinctes. Il y a donc une importante lacune dans la recherche sur les MpBC, que nous souhaitons combler avec la création d'un atlas transcriptomique complet, répertoriant l'ensemble des profils et signaux transcriptomiques de référence spécifiques pour chacun des sous-types tumoraux retrouvées dans les MpBC. Pour réaliser cela, nous prévoyons de compléter les données spatiales avec des données de scRNA-seq, permettant l'analyse ARN des cellules uniques pour chaque échantillon. Plus précisément cela correspondra à du single-nuclei RNAseq, (séquençage des noyaux uniques) car le protocole de fixation et d'inclusion en paraffine (FFPE) ne permet pas le séquençage des cellules dans son entièreté. Ces données transcriptomiques générés à l'échelle de la cellule unique permettront de créer un profil transcriptomique moyen, spécifique pour chaque sous-type cellulaire, auxquels nous pourrons nous rapporter lors de l'étape de déconvolution des spots. Des outils tel que RCTD marche très bien avec des données de transcriptomique spatiales comme nous les avons. Ainsi, à partir des données spatiales Visium déjà analysées, des annotations histologiques des spots Visium fournies par l'expert anatomopathologiste et de l'atlas obtenue via les nouvelles données transcriptomiques générées par snRNA-seq, il sera plus facile de définir un profil transcriptomique et un profil d'altération du nombre de copie spécifique à chaque sous-type tumoral. Les données seront moins bruitées, les réductions PCA captureront avec peu de composante potentiellement mieux la variance du jeu de données et la séparation entre compartiments transdifférenciés sera beaucoup plus clair sur la projection UMAP. Cela facilitera, la constitution des clusters et les marqueurs spécifiques de chaque cluster seront bien plus représentatifs des sous-types tumoraux. En conclusion, cet atlas transcriptomique dédié aux MpBC à résolution single-cell et spatiales permettra de décrypter l'hétérogénéité des cellules tumorales et de mieux comprendre les différents types de transdifférenciation.

#### 4.2.4 Microdissection des compartiments

##### Déterminants génétiques

**Mutations drivers divergentes et validation des CNA** Dans le but d'explorer en profondeur les dynamiques évolutives des échantillons biphasiques MpBC, des microdissections par capture laser vont être procédées afin de déterminer les caractéristiques génétiques uniques des deux compartiments de chaque échantillon. L'ADN spécifique à chaque compartiment pourra être extrait et les déterminants génétiques somatiques, telles que les mutations ponctuelles et les CNA seront analysées par séquençage. On pourra ainsi valider plus précisément les CNA identifiés avec mon analyse préliminaire. A noter que pour réaliser cette analyse des altérations génomiques, il est nécessaire d'avoir un tissu contrôle au sein de chaque échantillon, pour pouvoir comparer les altérations du tissu tumorale avec un tissu de référence. Or, certains échantillons MpBC ne présentent pas de tissu sain, nous utiliserons donc au sein de chaque échantillon le tissu du compartiment tumoral appariés comme référence. Par exemple, pour un échantillon MpBC présentant 2 compartiments transdifférenciés (l'un épithéliale et l'autre chondroïde), pour analyser les mutations génétiques du compartiment chondroïde, nous utiliserons le compartiment épithélial comme référence, et vice versa. Les mutations identifiées par cette analyse génétique

seront considérées comme spécifique d'un sous-type tumoral si celle-ci est complètement absente des autres sous-types. Enfin, en croisant les mutations drivers contenues dans différentes bases de données (COSMIC, TGCA) avec nos mutations identifiées, nous serons en mesure de préciser si la mutation est dite « clonales », c'est-à-dire une mutation présente dans la cellule cancéreuse à l'origine de tous les différents sous-types tumoraux.

**Epigénétique (méthylome)** Les microdissections de chaque compartiment tumoral pourront également permettre une analyse épigénétique de chaque sous-type. Une analyse du méthylome complétera nos données génétiques pour chaque sous-type, et permettre l'identification de modifications réversibles de gène soumis à des régulations provenant de l'environnement tumoral. Nous pourrons révéler les gènes suppresseurs de tumeurs hyperméthylés et inactivé par silencing, ou au contraire des oncogènes hypométhylés favorisant la prolifération et la survie, et l'instabilité génomique des cellules tumorales. Les données obtenues pourront être croisés avec les modifications épigénétiques connues et répertoriées dans la base de données (.). Ce type d'analyse est particulièrement intéressante dans notre projet car, il a été montré à de nombreuses reprises dans la littérature scientifique que certaines modifications épigénétiques sont à l'origine d'un remodelage de l'identité des cellules tumorales. Les cellules cancéreuses peuvent transitionner par des stades cellulaires immatures disposant d'une grande plasticité (proche des cellules souches) et acquérir des phénotypes très divers. (ref)

#### 4.2.5 Déterminants non génétiques

**Déconvolution et composition du microenvironnement & Interactions ligand-récepteur tumeur/TME et cœur/bordure** De plus, il est aujourd'hui connu que de nombreuses cellules immunitaires (lymphocytes T CD4, CD8, macrophages) et non immunitaires (CAF (Cancer Associated Fibroblast)) jouent un rôle prépondérant dans la constitution de niche biologique au sein du tissu tumoral, et régulant très finement la trajectoire évolutive des populations clonales tumorales. En effet, grâce à la déconvolution des spots Visium, nous pourront ainsi évaluer la composition du microenvironnement tumoral pour chaque compartiments transdifférenciés et déterminer si des facteurs externes pourraient être à l'origine de la transdifférenciation en un sous-type spécifique. L'identification de paires ligand-récepteurs permettra de révéler de potentielles interactions entre les cellules transdifférenciés et les cellules du micro-environnement.

#### 4.2.6 Validation IHC & Analyses fonctionnelles

Chacun des biomarqueurs identifiés feront l'objet d'une validation Immunohistochimique (IHC) pour chaque sous-type tumoral afin d'estimer leur pertinence dans le diagnostic des MpBC. Enfin, la culture d'organoide dérivé de patiente atteintes de MpBC, et l'utilisation de xénogreffes murines permettront d'étudier via des perturbation génétiques et pharmacologiques, l'impact fonctionnel de ces biomarqueurs identifiés. Afin de prévenir tout biais dans la validation et de garantir l'authenticité des résultats, ces analyses se feront à partir d'un panel d'échantillon indépendant de ceux utilisé pour déterminer ces marqueurs dans notre projet.

## 5 Conclusions

Les objectifs de ce stage étaient, dans un premiers temps, d'explorer les données transcriptomiques spatiales des carcinomes mammaires métaplasiques générées avec la technologie Visium. Cette première approche visait notamment à définir des gènes impliqués dans la transdifférenciation, et spécifiques de chaque sous-type tumoraux. Puis, via des analyses d'altération du nombre de copie, d'identifier des déterminants génétiques qui pourraient être à l'origine des transdifférenciations entre les compartiments tumorales.

A travers l'analyse de 16 premiers échantillons MpBC, j'ai pu identifier certains marqueurs phénotypiques comme EPCAM ou CD26 qui semblent être des gènes, a priori, spécifiques aux sous-types épithéliales des MpBC et retrouvé dans plusieurs des patients partageant ce type de transdifférenciation. Ces marqueurs nous donne des premières pistes à approfondir pour les analyses futurs qui permettront de réellement définir la spécificité et la représentativité des marqueurs phénotypiques de chaque sous-type tumoraux. De plus, mon travail a mis en évidence des altérations du nombre de copie localisés à des régions précises du génome, comme le bras 13q et Xq, retrouvées fréquemment délétées dans les cellules épithéliales.

Bien que certaines limitations, liées à la résolution spatiale des spots de la technologie Visium, empêche une réelle analyse approfondie et une identification claire du profil transcriptomique des sous-types, ces premiers résultats constituent tout de même une base solide et encourageante pour la suite des analyses nous permettant d'orienter nos prochaines recherches.

Ce travail m'a également permis de consolider grandement mes connaissances dans les analyses omics, en manipulant des données génomiques, transcriptomiques, et d'utiliser des outils complexes tel que Seurat et InferCNVPlus.

## Références

- [1] Organisation Mondiale de la Santé. Cancer du sein. <https://www.who.int/fr/news-room/fact-sheets/detail/breast-cancer>, 2022. Consulté le 18 mai 2025.
- [2] Fondation ARC pour la recherche sur le cancer. L'essentiel sur les cancers du sein. <https://www.cancer.fr/personnes-malades/les-cancers/sein/comprendre-les-cancers-du-sein/l-essentiel>, 2023. Consulté le 18 mai 2025.
- [3] Société canadienne du cancer. Le cancer du sein triple négatif. <https://cancer.ca/fr/cancer-information/cancer-types/breast/what-is-breast-cancer/cancerous-tumours/triple-negative-breast-cancer>, 2023. Consulté le 18 mai 2025.
- [4] BRCA France. Le cancer héréditaire du sein et de l'ovaire. <https://www.brcafrance.fr/cancer-hereditaire-sein-et-ovaire/>, 2022. Consulté le 18 mai 2025.
- [5] American Cancer Society. Triple-negative breast cancer. <https://www.cancer.org/cancer/breast-cancer/about/types-of-breast-cancer/triple-negative.html>, 2023. Consulté le 18 mai 2025.
- [6] Société canadienne du cancer. Grade et stade du cancer. <https://cancer.ca/fr/cancer-information/what-is-cancer/stage-and-grade/grading>, 2023. Consulté le 18 mai 2025.
- [7] Giampaolo Bianchini, Justin M Balko, Ingrid A Mayer, Melinda E Sanders, and Luca Gianni. Triple-negative breast cancer : challenges and solutions. *Nature Reviews Clinical Oncology*, 13(11) :674–690, 2016.
- [8] A. Coutant, V. Cockenpot, L. Muller, C. Degletagne, R. Pommier, L. Tonon, M. Ardin, M. C. Michallet, C. Caux, M. Laurent, A. P. Morel, P. Saintigny, A. Puisieux, M. Ouzounova, and P. Martinez. Spatial transcriptomics reveal pitfalls and opportunities for the detection of rare high-plasticity breast cancer subtypes. *Laboratory Investigation*, 103(12) :100258, 2023. Epub 2023 Oct 7.
- [9] CNRS. Est-ce qu'une division change les étapes de reprogrammation pour une cellule ? <https://www.insb.cnrs.fr/fr/cnrsinfo/est-ce-qu'une-division-change-les-etapes-de-reprogrammation-pour-une-cellule>, 2024. Consulté le 18 mai 2025.
- [10] Centre Léon Bérard. Le centre de ressources biologiques du clb. <https://www.centreleonberard.fr/professionnel-de-sante-chercheur/recherche-contre-le-cancer/recherche-translationnelle/le-centre-de-ressources-biologiques>, 2023. Consulté le 18 mai 2025.
- [11] Leidamarie Tirado-Lee. Spatially resolved transcriptomics : An introductory overview of spatial gene expression profiling methods. <https://www.10xgenomics.com/blog/spatially-resolved-transcriptomics-an-introductory-overview-of-spatial-gene-expression>, 2023. Consulté le 18 mai 2025.
- [12] 10x Genomics. Space ranger software (version 2.0.0). <https://www.10xgenomics.com/products/spatial-gene-expression>, 2023. Consulté le 18 mai 2025.

- [13] 10x Genomics. Loupe browser software (version 8.1.2). <https://www.10xgenomics.com/products/loupe-browser>, 2024. Consulté le 18 mai 2025.
- [14] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck, S. Zheng, Andrew Butler, M. J. Lee, Aaron J. Wilk, Chloe Darby, Michael Zager, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexis, Eleni Mimitou, and Rahul Satija. Integrated analysis of multimodal single-cell data. <https://satijalab.org/seurat/>, 2021. Seurat R package, version 5.1.0, consulté le 18 mai 2025.
- [15] Ilya Korsunsky, Nathan Millard, Jian Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yury Baglaenko, Michael Brenner, Po-Ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature Methods*, 16 :1289–1296, 2019. Package R Harmony, version 1.2.3, consulté le 18 mai 2025.
- [16] Charlene Zhang. infercnvplus : Enhanced 'infercnv' package. <https://github.com/CharleneZ95/infercnvPlus>, 2020. Consulté le 18 mai 2025.
- [17] Broad Institute. infercnv of the trinity ctat project. <https://github.com/broadinstitute/infercnv>, 2024. Consulté le 18 mai 2025, version stable recommandée par le dépôt.
- [18] UCSC Genome Browser. Ucsc genome browser : Human genome assembly grch38/hg38. <https://genome.ucsc.edu/cgi-bin/hgGateway?db=hg38>, 2013. Consulté le 18 mai 2025.
- [19] RStudio Team. Rstudio : Integrated development environment for r (version 2024.09.0+375 "cranberry hibiscus"). <https://posit.co/download/rstudio-desktop/>, 2024. Consulté le 18 mai 2025.
- [20] BRIC Bordeaux. Prix ruban rose avenir – dr monica arnedos, team 7 (bric). <https://www.bricbordeaux.com/en/2025/02/prix-ruban-rose-avenir-dr-monica-arnedos-team-7-bric/>, 2025. Consulté le 18 mai 2025.