



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 : Transcriptomique spatiale et évolution des carcinomes mammaires métaplasiques

---

**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 carcinome mammaire métaplasique (MpBC). Ce projet implique plusieurs centres de recherche au niveau national, et vise à combler nos connaissances encore incomplètes sur la biologie des MpBC, en particulier la transdifférenciation qui les définit. Cela se traduit par la présence de compartiments tumoraux non-épithéliaux rares, et les patientes sont confrontées 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 diagnostiquer moléculairement les différents sous-types de MpBC. De plus, j'ai pu mettre en évidence des régions chromosomiques significativement altérées entre des compartiments appariés mais de sous-type différent chez certaines patientes. Ces résultats contribueront à des analyses approfondies qui seront réalisées prochainement, pour déterminer les déterminants génétiques et non-génétiques de la transdifférenciation dans les MpBC. A plus long terme, ces travaux ouvriront la voie vers l'identification de marqueurs moléculaires pour le diagnostic, et de cibles thérapeutiques pour la prise en charge des patientes atteintes de ces cancers du sein rares et atypiques.

# Table des matières

|   |           |
|---|-----------|
| <b>Liste des abréviations</b>   | <b>3</b>  |
| <b>Liste des logiciels utilisés</b>   | <b>4</b>  |
| <b>1 Introduction</b>   | <b>5</b>  |
| 1.1 Contexte et état de l'art . . . . .   | 5         |
| 1.2 Présentation des MpBC . . . . .   | 5         |
| 1.3 Problématique(s) . . . . .  | 6         |
| 1.4 Pertinence de la transcriptomique spatiale . . . . .                                | 7         |
| <b>2 Matériels et méthodes</b>  | <b>8</b>  |
| 2.1 Echantillons MpBC . . . . .   | 8         |
| 2.1.1 Cohorte : CLB . . . . .   | 8         |
| 2.1.2 FFPE fixation et H&E coloration . . . . .   | 8         |
| 2.1.3 Séquençage Visium & alignement . . . . .  | 8         |
| 2.2 Annotation phénotypique des spots . . . . .   | 9         |
| 2.3 Données scRNA-seq . . . . .   | 9         |
| 2.3.1 Contrôle qualité et filtrage des spots . . . . .                                  | 9         |
| 2.3.2 Normalisation et Scaling . . . . .  | 10        |
| 2.4 Analyse des marqueurs phénotypiques . . . . .                                       | 10        |
| 2.4.1 Harmony . . . . .   | 10        |
| 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 langage 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 . . . . .                                 | 16        |
| 3.1.3 Identification des gènes marqueurs . . . . .                                      | 17        |
| 3.2 Analyse des altérations génomiques divergentes . . . . .                            | 19        |
| 3.2.1 Profils génomiques biphasiques . . . . .  | 19        |
| 3.2.2 Identification des altérations divergentes entre compartiments appariés . . . . . | 20        |
| <b>4 Discussion</b>   | <b>23</b> |
| <b>5 Conclusions</b>  | <b>27</b> |

## Table des figures

|   |   |    |
|---|---|----|
| 1 | <b>Figure 1 : Exemple d'échantillon MpBC mixte.</b> . . . . .   | 6  |
| 2 | <b>Figure 2 : Correction de l'effet <i>batch</i> entre les 16 échantillons distincts de MpBC.</b> . . . . .   | 16 |
| 3 | <b>Figure 3 : Projection UMAP des spots Visium.</b> . . . . .   | 17 |
| 4 | <b>Figure 4 : Visualisation de l'expression génique par cluster sur la projection UMAP.</b> . . . . .   | 18 |
| 5 | <b>Figure 5 : Profil d'expression des marqueurs par archétypes et par patients.</b> . . . . .   | 18 |
| 6 | <b>Figure 6 : Heatmap des résultats InferCNVPlus pour toutes les patientes.</b> . . . . .   | 19 |
| 7 | <b>Figure 7 : Représentation des scores CNA obtenus avec InferCNV-Plus, avant (A) et après normalisation (B), en fonction des cytobandes mineures pour chaque chromosome du génome.</b> . . . . . | 20 |
| 8 | <b>Figure 8 : Volcano Plot des altérations CNA divergentes entre compartiments tumoraux appariés.</b> . . . . .   | 21 |
| 9 | <b>Figure 9 : Analyse de la corrélation des scores CNA entre les deux sous-types tumoraux du patient MpBC9 et MpBC15.</b> . . . . .   | 22 |

## Liste des tableaux

|   |   |    |
|---|---|----|
| 1 | <b>Table 1 : Mesures de contrôle qualité de séquençage des 16 échantillons MpBC.</b> . . . . .                                      | 9  |
| 2 | <b>Table 2 : Tableau récapitulatif des différents compartiments tumoraux comparés dans l'analyse des CNA par patient.</b> . . . . . | 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

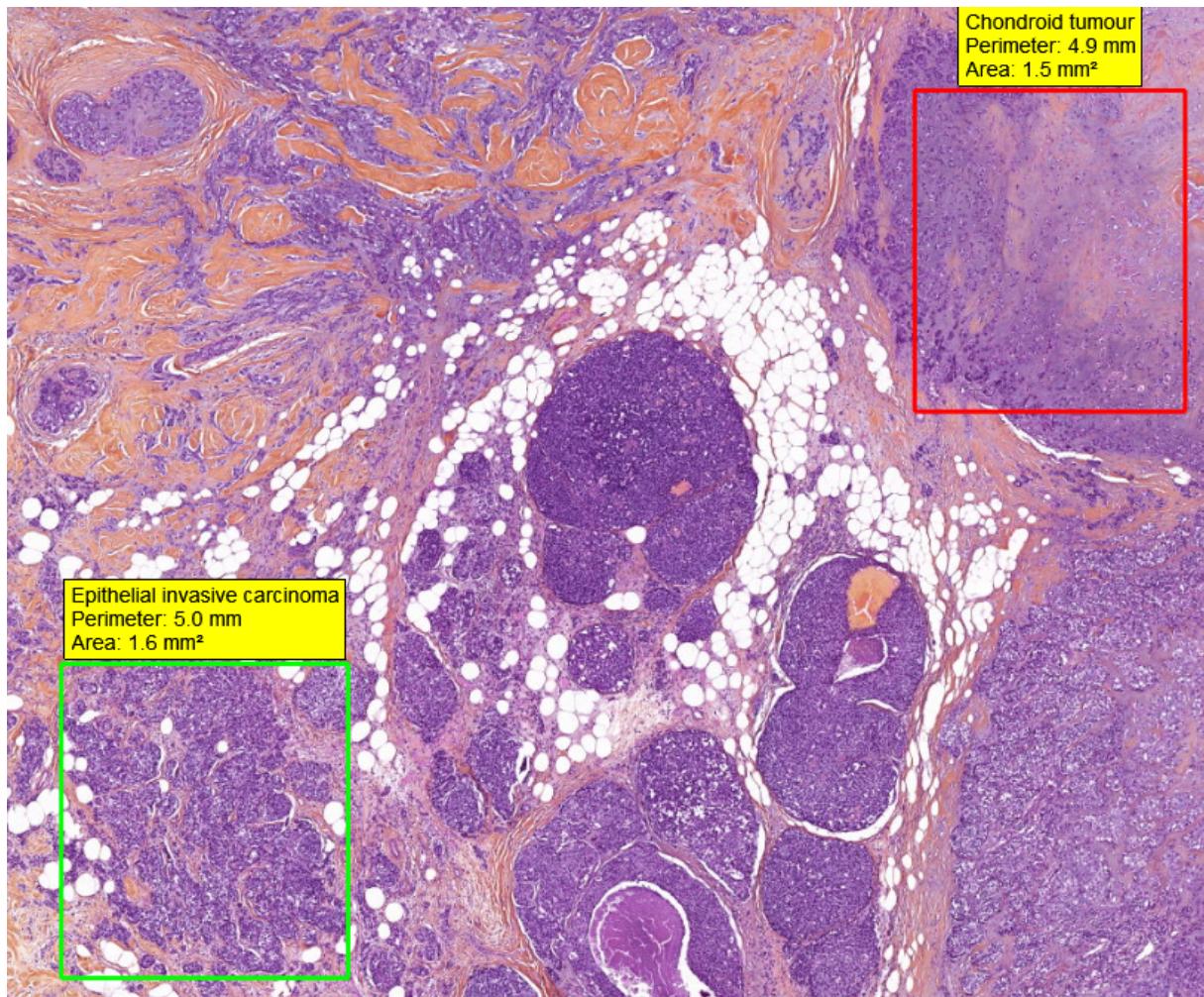
## 1.1 Contexte et état de l'art

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]. Cette pathologie est donc un enjeu majeur de santé publique. Parmi les cancers du sein, on distingue les types TNBC (Triple Negative Breast Cancer), caractérisés par l'absence de récepteurs aux œstrogènes et progestérone, et l'absence de surexpression du gène HER2. Ces absences empêchent ainsi les patientes de pouvoir répondre aux traitements par thérapies ciblées existantes. Ainsi, les TNBC sont la plupart du temps des tumeurs agressives et sont associées à une plus forte mortalité. Elles disposent également d'une capacité à se développer et à se propager très rapidement [2] [3]. Enfin, ces TNBC sont caractérisés par une grande variété de sous-types, avec des phénotypes très hétérogènes [4]. Parmi eux, nous retrouvons les carcinomes du sein métaplasiques (Metaplastic Breast Cancer, MpBC) [5], qui 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 agressive et avec une forte mortalité.

## 1.2 Présentation des MpBC

Dans ce projet de recherche nous nous intéressons plus spécifiquement aux MpBC, afin de mieux comprendre les mécanismes de plasticité cellulaire qui les caractérisent et compliquent leur prise en charge clinique. Les MpBC sont définis par la présence d'un compartiment tumoral transdifférencié, défini exclusivement de manière histologique par une pathologiste via la présence de cellules tumorales de type non-épithéial. Il n'existe à ce jour, pas de moyen moléculaire (analyses ADN ou ARN de type « omique », par exemple) de diagnostiquer les MpBC. Selon le CNRS, la transdifférenciation est « la conversion d'un type cellulaire entièrement différencié en un autre type » [6]. Les échantillons MpBC présentant au moins 2 compartiments tumoraux de type différent au sein de la tumeur sont dits « mixtes », en opposition aux MpBC « purs » ne présentant que des cellules cancéreuses transdifférenciées. Actuellement, plusieurs types de transdifférenciation des MpBC à partir des cellules tumorales épithéliales peuvent être observés, en particulier : Malpighienne (Squamous), Fusiforme (Spindle cell), Chondroïde et Osteosarcomatoïde. Selon le type de la transdifférenciation, les cellules tumorales peuvent par exemple présenter un phénotype épithéial épidermoïde différent des cellules cancéreuses initiales (transdifférenciation malpighienne), ou au contraire présenter un phénotype plus similaire au tissu stromal de soutien (transdifférenciation mésenchymateuse). Certaines cellules, lors de ce processus peuvent également prendre une forme allongée, en forme d'épingle (transdifférenciation fusiforme), ou bien créer abondamment de la matrice extracellulaire cartilagineuse (transdifférenciation chondroïde), voire même une matrice ostéoïde ressemblant à de l'os immature (transdifférenciation 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.** On observe la présence de deux compartiments tumoraux distincts au sein de la même tumeur : des cellules tumorales chondroïdes (en rouge) et des cellules tumorales épithéliales (en vert).

### 1.3 Problématique(s)

Dans ce projet de recherche, j'ai cherché à mieux comprendre l'origine de l'apparition de ces différents compartiments cellulaires au sein des MpBC et les mécanismes la sous-tendant, grâce à des données de transcriptomique spatiale sur des échantillons de MpBC mixtes. Mon travail s'est articulé autour de 2 principaux objectifs. Dans un premier temps, 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 cette transition entre deux phénotypes, en effectuant une analyse des altérations du nombre de copie (CNA) les deux compartiments de chaque MpBC mixte.

## 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 mixtes de patientes. Cette technologie à résolution spatiale est particulièrement intéressante dans ce projet de recherche car elle va permettre d'associer les annotations pathologiques des compartiments, réalisées par une experte anatomopathologiste, avec des analyses transcriptomiques. La nature même des MpBC mixtes fait qu'il existe plusieurs compartiments cellulaires au sein même de la tumeur. L'utilisation de technologies *bulk*, telles que le -RNAseq, serait non pertinente dans ce contexte, car nous perdrions l'information spatiale et le signal de chaque compartiment serait mélangé en un ensemble non-spécifique. Ainsi, avec une information spatiale en plus du profil transcriptomique, il est possible d'étudier chaque compartiment de manière individuelle avec une très haute résolution ( 10 cellules par « spot »). Cela permettra d'établir des marqueurs moléculaires spécifiques à chaque sous-type cellulaire transdifférencié, qui pourront être utilisés par la suite pour le diagnostic et l'identification de potentielles cibles thérapeutiques. Enfin, cela permettra des analyses génomiques à faible résolution, mais spécifique à chaque compartiment.

## 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 été 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 assurées par la Plateforme de Gestion des Echantillons Biologiques (PGEB) du CLB [7]. 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 conservée 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, permet 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 segmentée 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) [8], ou *barcode*, puis une amplification de ces ARN par PCR [9]. 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.

Le Table 1 ci-dessous répertorie toutes les mesures de contrôle qualité de séquençage pour les 16 échantillons MpBC.

TABLE 1 – Mesures de contrôle qualité de séquençage des 16 échantillons MpBC.

| Sample ID | Visium Slide Number | Sequencing Batch | Tumoral Compartment 1 | Tumoral Compartment 2  | Spot Under Tissue | Reads Mapped (%) | Mean Reads Per Spot | Median Genes Per Spot | Valid Barcode (%) | Sequencing Saturation (%) |
|-----------|---------------------|------------------|-----------------------|------------------------|-------------------|------------------|---------------------|-----------------------|-------------------|---------------------------|
| MpBC1     | 1                   | 1                | Epithelial tumor      | Mesenchymal tumor      | 4992              | 96.4             | 65225               | 4982                  | 97.9              | 81.1                      |
| MpBC2     | 1                   | 1                | Epithelial tumor      | Spindle-cell tumor     | 4992              | 99.0             | 61529               | 8181                  | 99.2              | 39.1                      |
| MpBC3     | 2                   | 1                | Epidermoid tumor      | Spindle-cell tumor     | 4992              | 99.0             | 64519               | 4480                  | 99.1              | 82.7                      |
| MpBC4     | 2                   | 1                | Spindle-cell tumor    | Osteosarcomatoid tumor | 4992              | 98.9             | 72583               | 6565                  | 99.1              | 68.8                      |
| MpBC5     | 3                   | 1                | Epithelial tumor      | Spindle-cell tumor     | 4992              | 99.1             | 50449               | 6146                  | 99.2              | 44.7                      |
| MpBC6     | 3                   | 1                | Spindle-cell tumor    | NA                     | 4992              | 98.9             | 57511               | 5380                  | 99.1              | 76.6                      |
| MpBC7     | 4                   | 1                | Spindle-cell tumor    | NA                     | 4992              | 99.0             | 43589               | 3634                  | 99.1              | 77.6                      |
| MpBC8     | 4                   | 1                | Epidermoid tumor      | Mesenchymal tumor      | 4992              | 99.1             | 64339               | 7922                  | 99.2              | 46.8                      |
| MpBC9     | 5                   | 2                | Epithelial tumor      | Chondroid tumor        | 4992              | 97.6             | 19251               | 842                   | 98.7              | 91.9                      |
| MpBC10    | 5                   | 2                | Epithelial tumor      | Spindle-cell tumor     | 2589              | 97.9             | 34699               | 3144                  | 99.0              | 83.2                      |
| MpBC11    | 6                   | 2                | Epithelial tumor      | Spindle-cell tumor     | 4992              | 89.5             | 8921                | 189                   | 91.1              | 96.3                      |
| MpBC12    | 6                   | 2                | Epithelial tumor      | Spindle-cell tumor     | 3881              | 63.6             | 8627                | 0                     | 24.6              | 99.6                      |
| MpBC13    | 7                   | 2                | Epithelial tumor      | Chondroid tumor        | 4992              | 97.6             | 9877                | 1136                  | 98.6              | 79.3                      |
| MpBC14    | 7                   | 2                | Epithelial tumor      | Spindle-cell tumor     | 4992              | 98.1             | 69479               | 6736                  | 98.8              | 65.7                      |
| MpBC15    | 8                   | 2                | Epithelial tumor      | Chondroid tumor        | 4992              | 98.0             | 15583               | 1506                  | 98.9              | 80.3                      |
| MpBC16    | 8                   | 2                | Epithelial tumor      | Chondroid tumor        | 4066              | 97.8             | 15499               | 1520                  | 98.8              | 78.2                      |

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, et le logiciel Space Ranger (version 2.0.0) [10].

## 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 observés comme majoritaires sur la lame appariée 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) [11] 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 pathologiste 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 filtrée selon certaines caractéristiques : le nombre d’UMI (Unique Molecular Identifier) compté par spot, et le nombre de gène différent compté par spot. Dans notre analyse, nous avons sélectionné uniquement les spots 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 ou des

cellules endommagées. Une limite supérieure correspondant 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** 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) [12]. 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 déterminé de façon empirique en fonction de la qualité des résultats obtenus en aval. Pour notre jeu de données, 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 de 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 réalisées 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) [13]. Après une réduction PCA de notre matrice de comptage globale sur 50 composantes, les variations non biologiques

ont été contrôlées 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 utilisée pour le séquençage (2 patientes différentes 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 réduit 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 une 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 dans l’espace de dimension réduit.

**Clustering et détermination des archétypes** 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 20 premières composantes principales corrigées par `Harmony` comme base du graphe. Puis nous avons appliqué un algorithme de clustering de type Multi-Louvain, regroupant les spots en cluster. 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 regardé la stabilité des clusters, le nombre de types cellulaires attendus (correspondant aux différents sous-types tumoraux) et en comparant les compositions en spot des clusters formés selon les types cellulaires annotés par l’experte pathologiste. La position relative des clusters sur la projection UMAP nous a permis d’identifier des regroupements de spot spécifiques à chaque sous-type tumoral. Un ou plusieurs clusters ont donc été associés à chaque annotation. Afin de construire les archétypes spécifiques à chaque sous-type, nous avons conservé uniquement les spots dont l’annotation fournit par la pathologiste concordait avec l’annotation attribuée aux clusters. Les spots présentant des discordances au niveau des annotations ont quant à eux été exclus (environ 10% des spots initiaux).

**Expression différentielle** Une analyse des gènes différemment exprimés (DGEA) a été réalisée pour chaque archétypes 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 différemment exprimés. Cette méthode est plus adaptée aux 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 exprimés 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 (c'est à dire le nombre de spots) entre les différents échantillons pour un sous-type de compartiment donné.

## 2.5 Analyse des altérations génétiques

### 2.5.1 InferCNVPlus

**Constitution du groupe de cellule de références** InferCNVplus [14], est une extension de la méthode InferCNV [15], et permet d'inférer les altérations du nombre de copies (CNA) 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. Cet outil prend en compte la position des gènes dans le génome et identifie les variations d'expression des régions chromosomiques correspondant à un profil d'altération CNA. A chaque région, l'outil attribue un score de CNA entre -1 (délétion maximale) et 1 (amplification maximale). Une région neutre sans CNA aura un score voisin de 0. 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 tous les 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 échantillon MpBC ont été comparé à un groupe de référence, commun à tous les échantillons. Les spots considérés comme non tumoraux étaient ceux annotés comme contenant 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 [16]). 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. Cela a permis de réduire le nombre de données CNA, tout en restant représentatif de la majorité des données. De plus, en lissant le signal CNA par

réion chromosomique, cela nous a permis de baser le signal sur des régions génomiques composées de plusieurs gènes, et donc d'être moins sensibles aux biais techniques et biologiques (variation du nombre d'UMIs entre types cellulaires, effet « *burst* » de l'expression ARN...) [17]. 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 2 récapitule les comparaisons de score CNA des tissus tumoraux réalisées par patient.

TABLE 2 – Tableau récapitulatif des différents compartiments tumoraux comparés dans l'analyse des CNA par patient.

| Patients | Tumoral Sub-types 1 | Tumoral Sub-types 2 |
|----------|---------------------|---------------------|
| MpBC1    | Epithelial          | Mesenchymal         |
| MpBC2    | Epithelial          | Spindle-cell        |
| MpBC3    | Epidermoid          | Spindle-cell        |
| MpBC4    | Osteosarcomatoid    | Spindle-cell        |
| MpBC5    | Epithelial          | Spindle-cell        |
| MpBC8    | Epidermoid          | Mesenchymal         |
| MpBC9    | Epithelial          | Chondroid           |
| MpBC10   | Epithelial          | Spindle-cell        |
| MpBC11   | Epithelial          | Spindle-cell        |
| MpBC13   | Epithelial          | Chondroid           |
| MpBC14   | Epithelial          | Spindle-cell        |
| MpBC15   | Epithelial          | Chondroid           |
| MpBC16   | Epithelial          | Chondroid           |

**Normalisation** Les spots de la technologie Visium correspondent à environ 55µm 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 les 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 InferCNVPlus. 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 (score CNA < 0) et d'amplification (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 épithéial). Enfin, par souci de cohérence pour la visualisation, tous les scores CNA par cytobande ont été ramenés

à une échelle entre -1 et 1, comme c'est le cas par gène.

### 2.5.2 Définition des altérations génomiques divergentes

**Tests statistiques et VolcanoPlot** Enfin, nous avons défini, pour chaque bras chromosomique (p ou q), les cytobandes qui le composent. Afin de tester la significativité des altérations génomiques divergentes entre les différents compartiments tumoraux pour chaque patient, nous avons testé si la distribution des différences de score CNA de toutes les cytobandes mineures pour chaque bras chromosomique (par exemple le bras p du chromosome 1 : 1p) était significativement différente de celle 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 (Student's t test) était réalisé, et un test non paramétrique de Wilcoxon dans le cas contraire. Nous avons ainsi obtenu une p-value pour chaque bras chromosomique de chaque patiente. Une méthode de correction stricte, correspondant à la correction de Bonferroni a été appliquée sur chacune des p-value du test statistique afin de corriger les biais liés aux tests multiples. Les bras chromosomiques ont été considérés comme significativement différents entre compartiments si la p-value corrigée était inférieure à un seuil de 0.001. Un second seuil a été défini par bras chromosomique afin d'estimer l'amplitude (soit l'intensité) des altérations divergentes entre compartiments. Pour cela, nous avons calculé la différence absolue des scores CNA médians entre compartiment par bras chromosomique, puis fixé un seuil correspondant au 90e percentile de la distribution de ces valeurs. Ainsi, seules les 10% d'altérations présentant les plus fortes amplitudes (en valeur absolue) ont été identifiées comme fortement amplifiées ou délétées. Cela a permis d'éliminer les bras chromosomiques significatifs mais ayant une taille d'effet trop peu importante, et donc potentiellement moins représentatifs de réelles différences génétiques sous-clonales

## 2.6 Software et packages

### 2.6.1 RStudio et langage R

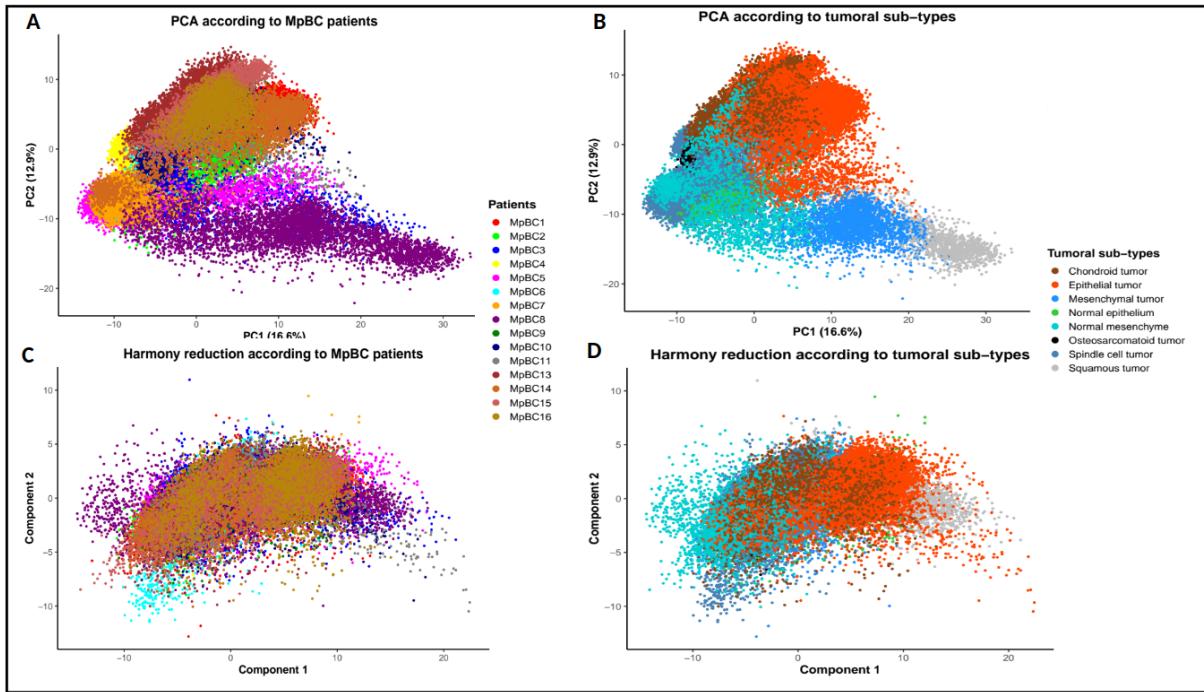
**Version du logiciel** L'ensemble des analyses bio-informatiques, ainsi que la génération des figures ont été réalisés à l'aide de scripts en R, développés dans l'IDE (Environnement de Développement Intégré) RStudio (version 2024.09.0+375 « Cranberry Hibiscus ») [18].

### 3 Résultats

#### 3.1 Analyse des marqueurs phénotypiques

##### 3.1.1 Réduction de dimensionalité

Une fois les étapes de contrôle qualité et normalisation effectuée, une première étape de réduction de dimension linéaire a été réalisée grâce à une ACP (Analyse en Composantes Principales) afin d'analyser le profil transcriptomique de chaque échantillon. La **figure 2A** représente les deux premiers axes de 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 données. La **figure 2B** représente les spots colorés selon les différents sous-types tumoraux. On constate qu'avec cette ACP, il est déjà possible de distinguer séparément les sous-types. Cela montre que nous capturons bien un signal biologique avec nos données. 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, ce qui peut biaiser les analyses en aval. Nous avons donc corrigé cet effet *batch*, dépendant du patient, avec la méthode Harmony [15]. On constate avec la **figure 2C** que les spots colorés, ne se séparent plus selon les patientes et se superposent. Cela suggère que la réduction de dimension corrigée par Harmony a bien fonctionné et permet donc d'atténuer la variation dues aux patientes d'origine, pour concentrer notre analyse sur les véritables différences biologiques. La **figure 2D** nous montre cette réduction de dimension linéaire corrigée par Harmony selon tous les sous-types tumoraux retrouvés chez nos 16 patientes. Bien que les biais techniques (*batch* de séquençage, lame utilisés, patiente d'origine) aient été corrigés, on s'aperçoit avec cette **figure 2D**, que les différences biologiques dans notre jeu de données (représenté par les sous-types tumoraux) sont toujours capturées. En effet, on constate clairement sur la première composante de cette réduction, un axe allant des types cellulaires épithéliaux (droite) aux types mésenchymateux (gauche). 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 données, en corrigeant les biais techniques tout en conservant les différences biologiques que l'on souhaite étudier.

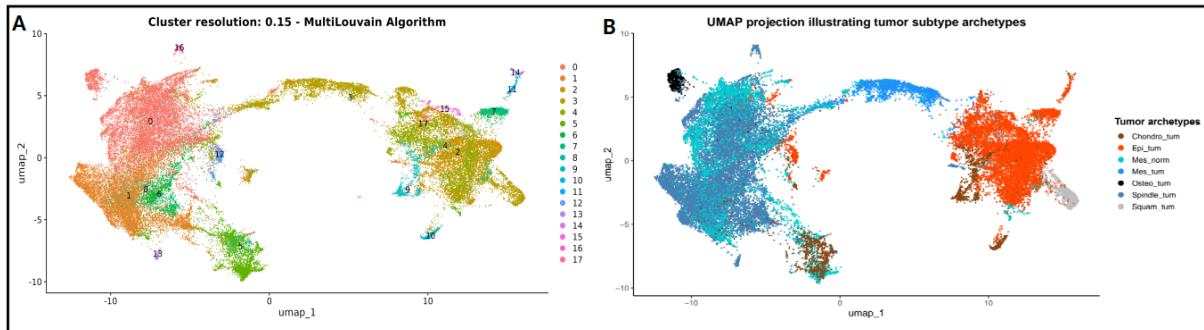


**FIGURE 2 – Correction de l’effet *batch* entre les 16 échantillons distincts de MpBC.** Analyse en composantes principales (ACP) des données transcriptomiques, selon les deux premières composantes. Avant correction par Harmony (*A - B*) et après correction (*C - D*), chaque spot est coloré soit selon le patient d’origine (*A, C*) soit selon les sous-types tumoraux (*B, D*).

### 3.1.2 Clustering et enrichissement archétypal

Nous avons utilisé les composantes principales corrigées des effets de *batch* issues de Harmony comme base de projection pour la visualisation UMAP. UMAP est ici préféré à t-SNE car il préserve mieux les structures globales tout en maintenant une bonne séparation des structures locales [19]. Le clustering Multi-Louvain a permis de déterminer les clusters de spots représentatifs des différentes annotations de sous-types tumoraux. Chaque annotation a ensuite été associée à regroupement de un ou plusieurs clusters. Les spots dont l’annotation ne concordait pas avec celle des regroupements de cluster ont été exclus (Voir Matériels et méthodes). La combinaison de ces informations a permis d’obtenir des « archétypes », soit les spots les plus représentatifs de chaque annotation des sous-types, illustrés en **figure 3**. On constate à nouveau une première composante discriminant les sous-types tumoraux selon un axe épithélio-mésenchymateux. On retrouve le l’archétype spécifique des cellules tumorales malpighiennes (Squam\_tum) avec le profil le plus épithéial, proche des cellules tumorales épithéliales (Epi\_tum). On peut également voir l’archétype spécifique aux ostéosarcomatoïdes bien distinct des autres archétypes. L’archétype spécifique des cellules fusiformes (Spindle\_tum) se positionnent à l’opposé des spots épithéliaux. Cela reflète bien la biologie observée des cellules tumorales dans les MpBC. Toutefois, on constate que les frontières de certains archétypes restent difficiles à caractériser et se fragmentent en plusieurs sous-clusters. C’est notamment le cas pour l’archétype spécifique des cellules tumorales fusiformes (Spindle\_tum) et tumorales chondroïdes (Chondro\_tum), qui sont plus difficiles à caractériser. Enfin, les spots annotés comme mésenchymateux (Mes\_tum) se positionnent entre les 2 gros regroupements de cellules considérées comme épithéliales et fusiformes, tandis que les spots de cellules mésenchymateuses normales se répartissent dans tous

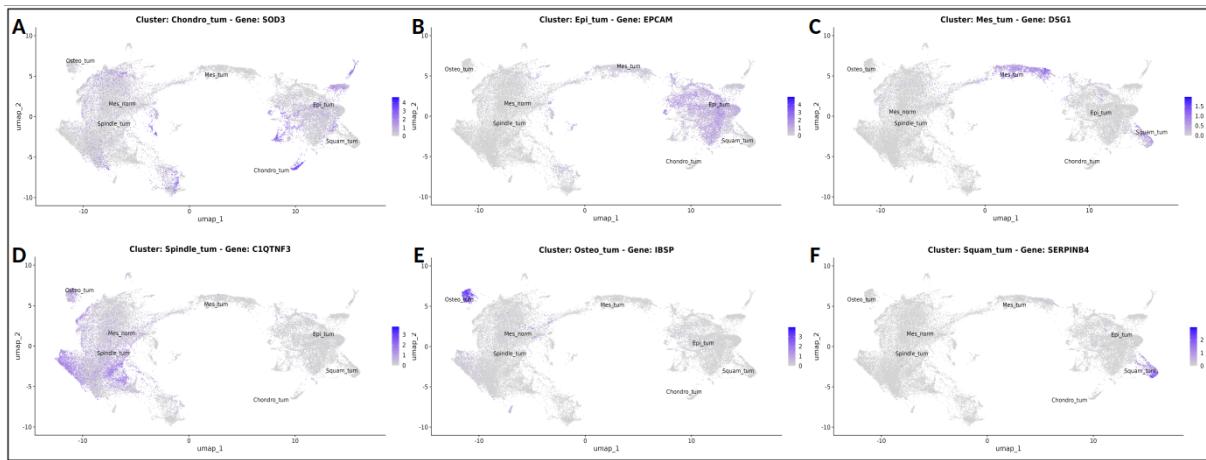
les archétypes. Cela est notamment dû aux limites de la technologie Visium, qui capture par plusieurs cellules spot et donc intègrent parfois des cellules minoritaires de types différents de l'annotation unique attribuée au spot. Pour conclure, cette projection UMAP faite à partir des données transcriptomiques des 16 patients MpBC a permis de caractériser 6 archétypes de cellules tumorales correspondant à 5 types de transdifférenciation en plus des cellules épithéliales typiques des TNBC. Ces archétypes ont par la suite été utilisés pour déterminer des marqueurs phénotypiques (gènes surexprimés) spécifiques à chacun d'eux.



**FIGURE 3 – Projection UMAP des spots Visium.** Projection UMAP des spots Visium et coloration selon les clusters identifiés via l'algorithme Multi-Louvain (**A**) et selon les archétypes spécifiques de chaque sous-types (**B**).

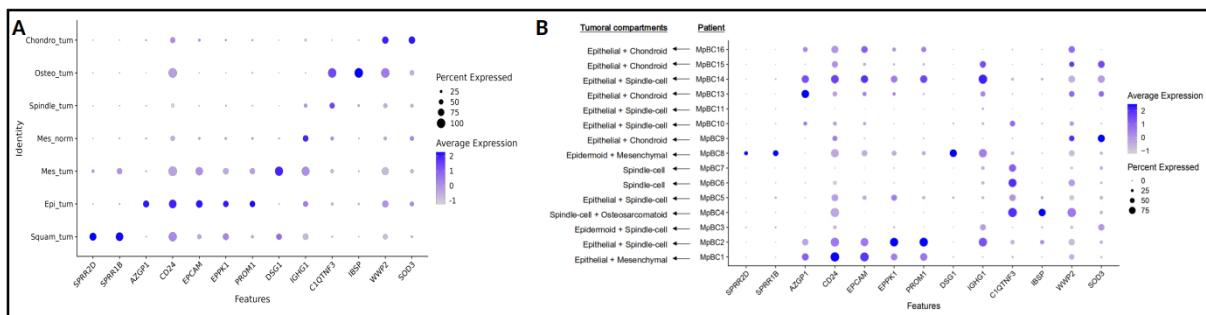
### 3.1.3 Identification des gènes marqueurs

A partir des archétypes identifiés précédemment, et représentant les différents sous-types tumoraux caractéristiques des MpBC, nous avons effectué une analyse des gènes différentiellement exprimés entre ces archétypes. En utilisant l'algorithme MAST, qui est un outil adapté au données single-cell, nous avons notamment identifié plusieurs marqueurs phénotypiques étant spécifiquement sur-exprimés dans chacun des archétypes. Les plus significatifs et biologiquement intéressants d'entre eux sont illustrés via la projection UMAP de la **figure 4**. Les cellules tumorales épithéliales (EPCAM, **Figure 4B**), ostéosarcomatoïdes (IBSP, **Figure 4E**) et malpighiennes (SERPINB4, **Figure 4F**) présentent des marqueurs à la fois spécifiquement surexprimés dans ces archétypes, représentatifs de tous les spots composant ces archétypes et cohérents avec leur phénotype. A l'inverse, les marqueurs identifiés pour les cellules tumorales chondroïdes (SOD3, **Figure 4A**), mésenchymateuses (DSG1, **Figure 4C**) et spindle-cell (C1QTNF, **Figure 4D**) manquent de spécificité (exprimés dans d'autres archétypes) ou de représentativité (expression dans seulement une fraction des spots archétypaux).



**FIGURE 4 – Visualisation de l’expression génique par cluster sur la projection UMAP.** La projection UMAP représente la distribution spatiale des archéotypes. Chaque panneau montre la surexpression d’un marqueur phénotypique au sein des cellules tumorales chondroïdes (A), tumorales épithéliales (B), tumorales mésenchymateuses (C), spindle-cell (D), ostéosarcomatoïdes (E), tumorales malpighiennes (F), selon une échelle de couleur (du gris pour une faible expression au bleu pour une expression élevée).

Afin de mieux caractériser l’expression de ces marqueurs archétypiques, et leur pertinence pour notre problématique, nous avons analysé l’expression moyenne et le pourcentage exprimé dans les spots de cellules des meilleurs marqueurs, selon chacun des archétypes identifiés (**figure 5A**) ou la patiente d’origine (**figure 5B**). Nous pouvons constater sur la **figure 5A** que certains de ces marqueurs, comme le gène EPCAM, sont surexprimés dans la majorité des patientes présentant le type cellulaire étudié et faiblement exprimés dans les autres archétypes. De la même façon, le gène SPRR1B semble être un marqueur pertinent pour caractériser le archétypes des cellules tumorales malpighienne (**figure 5A**). Cependant, en regardant la **figure 5B**, on constate que ce marqueur n’est exprimé uniquement pour une seule patiente (MpBC8). Or la patiente MPBC3 présente aussi ce sous-type tumoral malpighien, et on ne retrouve pas d’expression de ce marqueur pour cette patiente. Ainsi donc, cela nous montre qu’au-delà d’une analyse sur la totalité des cellules de chaque archétype, il sera nécessaire de réaliser une analyse plus poussée pour garantir que les marqueurs les plus significatifs ne soient pas patient-dépendants.

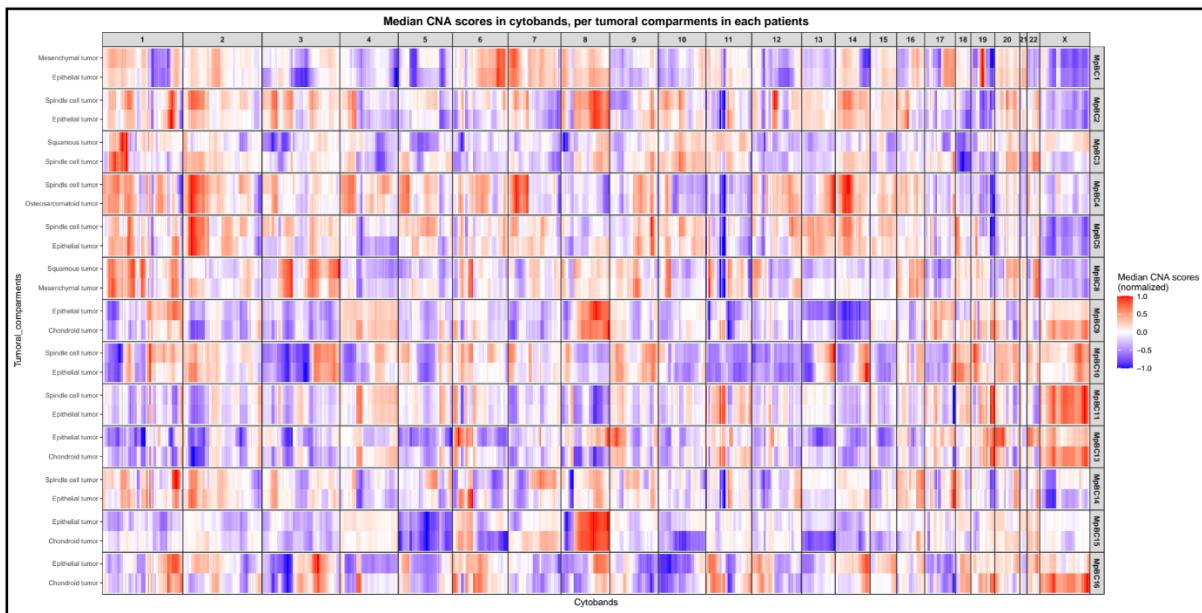


**FIGURE 5 – Profil d’expression des marqueurs par archétypes et par patients.** Représentation DotPlot de l’expression normalisée des gènes marqueurs sélectionnés selon les archétypes spécifiques de chaque sous-types tumoraux (**A**) ou selon les patientes atteintes de MpBC (**B**). La taille du point indique la proportion de cellules exprimant le gène, et la couleur du point reflète l’intensité moyenne d’expression normalisée.

## 3.2 Analyse des altérations génomiques divergentes

### 3.2.1 Profils génomiques biphasiques

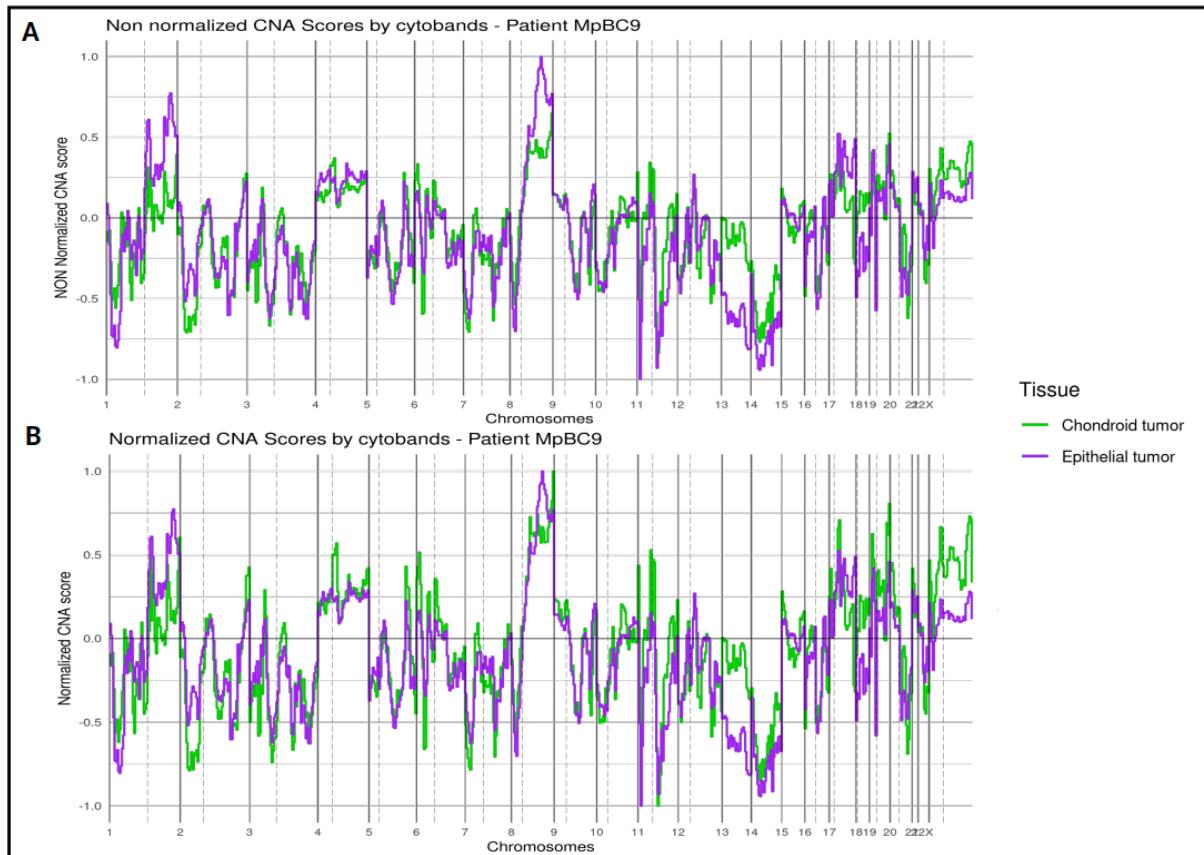
**Réduction par cytobande** Pour obtenir le profil de CNA (*Copy Number Alteration*) par patiente, nous avons utilisé le package R InferCNVPlus. Les résultats fournis par ce package se présentent sous la forme de matrices de scores CNA relatifs normalisés par échantillon : -1 pour la perte de matériel génétique maximale, +1 pour le gain maximal, et 0 pour le nombre de copies médian. Chaque ligne correspond à un spot et chaque colonne correspond à un gène. Dans notre analyse nous n'avons pas interprété directement les matrices brutes d'InferCNVPlus. Pour chaque patiente, nous avons plutôt analysé le profil CNA par compartiment tumoral (en faisant la médiane des scores par spot du même compartiment) et par cytobande (en faisant la médiane des scores par gène d'une même cytobande). Cela nous a permis d'obtenir la heatmap représenté en **figure 6**. On constate que les profils CNA des compartiments tumoraux appariés de chaque échantillon sont similaires pour la majorité des patientes et qu'on n'observe pas (ou peu) d'altérations du nombre de copie divergentes entre compartiment appariés.



**FIGURE 6 – Heatmap des résultats InferCNVPlus pour toutes les patientes.** Les scores CNA médian par cytobande sont montré pour chaque compartiment tumoraux au sein de chaque patiente. Les compartiments tumoraux appariés dans les échantillons MpBC présentent des profils CNA similaires.

**Normalisation intra-patient et inter-compartiments.** Nous travaillons ici avec des spots composés de plusieurs cellules, et donc avec une résolution supra-cellulaire. De plus il est important de noter que les spots peuvent présenter des profondeurs de séquençage différentes, en particulier selon le sous-type tumoral considéré. Nous avons donc procédé à une étape de normalisation afin de corriger les scores CNA des spots entre eux, à l'aide d'un facteur d'ajustement basé sur le rapport des scores CNA extrêmes entre les 2 tissus (voir Méthodes). Par exemple, pour la patiente MpBC9, le tissu tumoral épithéial présentait la plus forte amplification (**figure 7A**, chromosome 8), à laquelle le tissu tumoral chondroïde a été ajusté (**figure 7B**). La même correction a été appliquée pour les délétions (scores CNA < 0). Enfin, afin de faciliter la vi-

sualisation, les scores ont été renormalisés pour couvrir une échelle de -1 à 1, comme le signal InferCNVPlus initial. Cette normalisation a donc permis de faire coïncider les extrêmes des scores de délétion et d'amplification entre les différents compartiments tumoraux et de limiter les biais dus aux différences de profondeurs de séquençage et composition entre les différents compartiments. Par exemple, l'altération divergente du bras 8q avant normalisation (**figure 7A**) est complètement corrigée après ajustement (**figure 7B**). En revanche, d'autres différences, comme sur le bras 13q, persistent même après normalisation. En conclusion, cette normalisation a permis de corriger le signal de divergence entre CNA des compartiments appariés, pour garantir qu'ils n'étaient pas dûs à des différences d'abondance d'ARN ou de cellules.

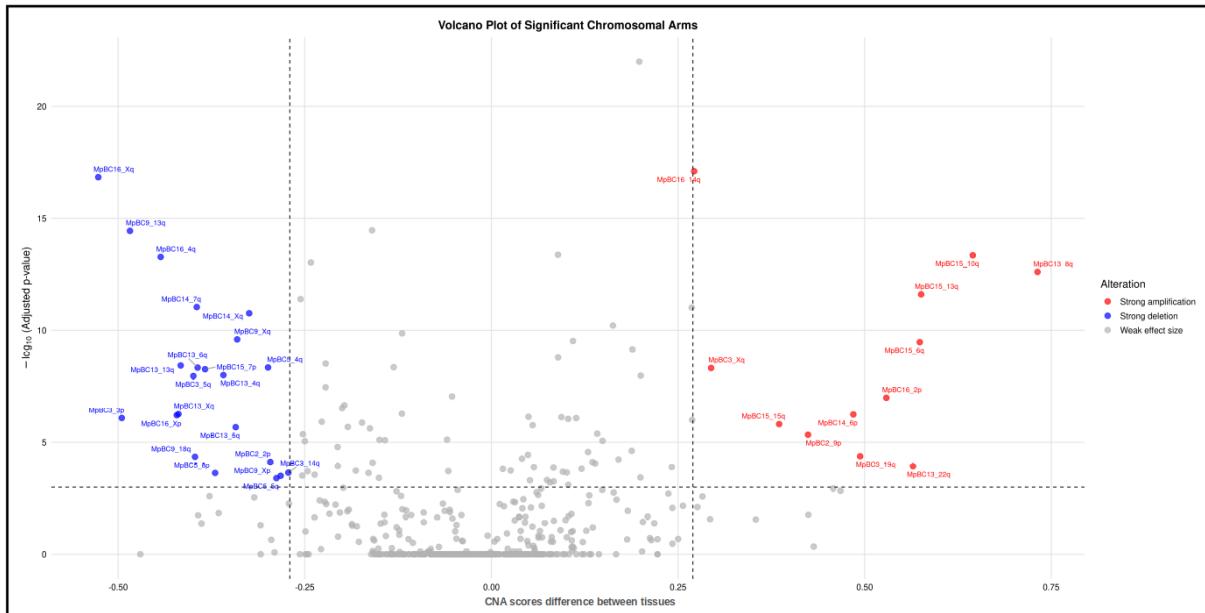


**FIGURE 7 – Représentation des scores CNA obtenus avec InferCNVPlus, avant (A) et après normalisation (B), en fonction des cytobandes mineures pour chaque chromosome du génome. La ligne verticale en pointillé indique la position du centromère pour chaque chromosome. Les couleurs des courbes correspondent aux deux sous-types tumoraux comparés chez le patient MpBC9 (chondroïde tumorale en violet et épithéial tumorale en vert). Les scores CNA s'étendent de -1 à 0 (délétion) et de 0 à 1 (amplification).**

### 3.2.2 Identification des altérations divergentes entre compartiments appariés

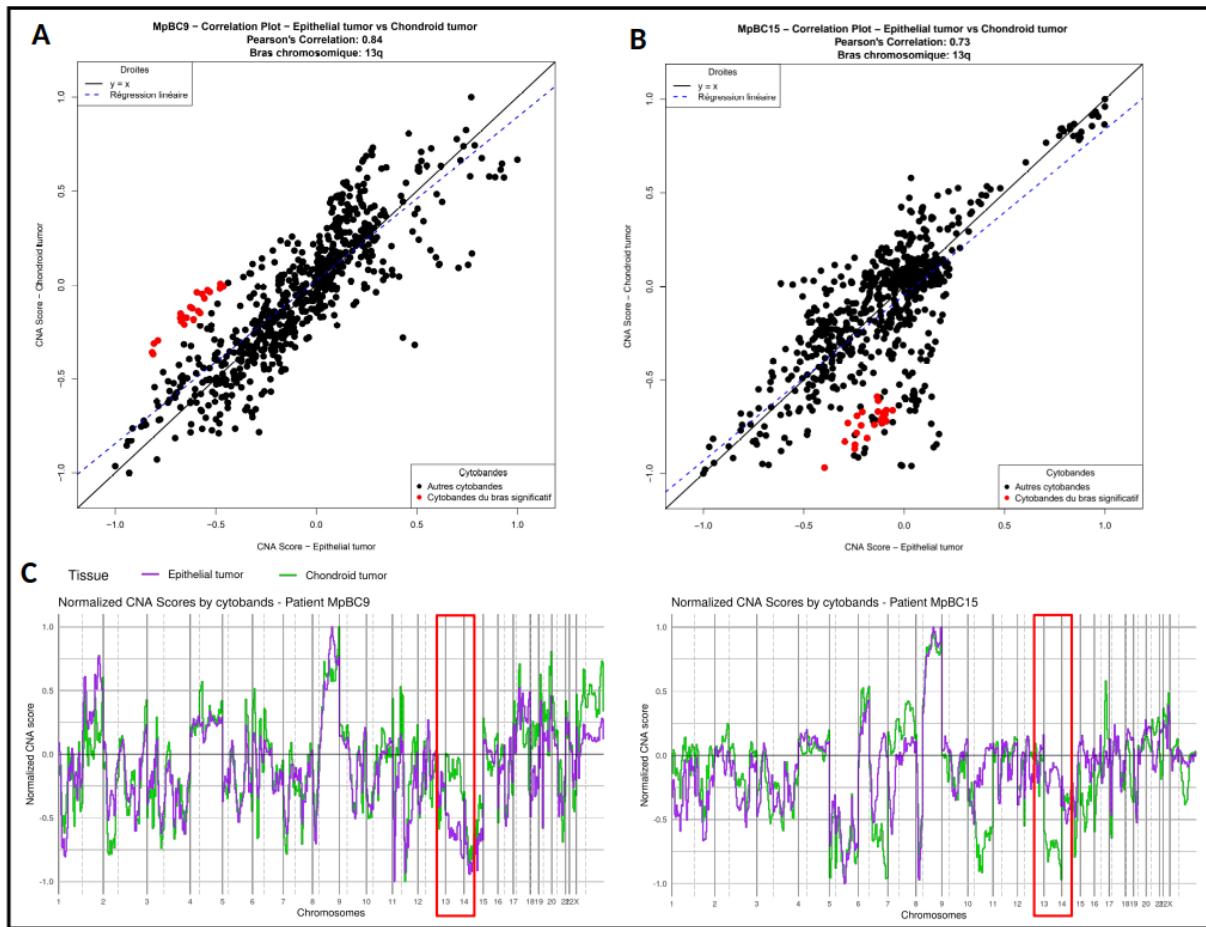
**Test par bras chromosomique** Afin d'identifier les altérations génomiques divergentes entre les compartiments appariés de chaque tumeur, nous avons testé si la distribution des différences de scores CNA entre compartiments pour un bras chromosomique donné, était significativement différente des différences dans les autres cytobandes. Les p-values obtenues lors des tests ont été corrigées, et considérées comme significatives lorsque la probabilité de commettre une erreur était très faible ( $\alpha=0.001$ ). En plus de ce seuil alpha, une deuxième métrique plus représentative la taille de l'effet a été utilisée, à savoir la différence médiane entre les compartiments

pour chaque bras chromosomique (voir Méthodes). La **figure 8** permet de représenter les bras chromosomiques significatifs par patiente sous forme d'un VolcanoPlot. La droite horizontale en pointillée indique le seuil de significativité ( $\alpha=0.001$ ), tandis que les droites verticales en pointillés délimitent les seuils définissant une forte différence absolue de score entre compartiments. Par exemple, le bras chromosomique le plus significativement délété et avec la plus grande différence entre les compartiments tumoraux correspond au bras Xq chez la patiente MpBC16. Toutefois, bien que certains bras chromosomiques présentent des altérations divergentes significatives entre compartiment tumoraux, nous pouvons voir que cela ne concerne que quelques patientes uniquement.



**FIGURE 8 – Volcano Plot des altérations CNA divergentes entre compartiments tumoraux appariés.**  
Pour chaque patiente, les bras chromosomiques sont représentés selon la p-value ajustée (correction de Bonferroni, seuil  $\alpha = 0,001$ ) et la différence absolue de score CNA entre compartiments (seuil  $\geq 0,27$ ). Seules quelques patientes présentent des altérations significatives.

Enfin, afin de d'investiguer plus en détail les bras chromosomiques significativement divergents, nous avons représenté la corrélation des scores CNA par cytobandes pour chacun des compartiments tumoraux appariés de chaque échantillon. Un exemple des résultats pour les patientes MpBC9 et MpBC15 sont représentés sur la **Figure 9**. Chaque point représente le score normalisé obtenu pour chaque cytobande. Les cytobandes appartenant au bras retrouvé comme significatif par le test statistique ont été colorées en rouge. On constate que pour le bras chromosomique 13q chez MpBC9 (**figure 9A**) et MpBC15 (**figure 9B**), les cytobandes appartenant à ces bras sont outliers par rapport aux autres cytobandes. La **figure 9C** illustre ces altérations divergentes significatives sur le profil CNA. L'encadré rouge montre l'emplacement du bras chromosomique significativement divergent au sein du profil CNA. Pour conclure, cette figure montre que nous retrouvons quelques rares cas d'altération divergentes entre les compartiments tumoraux appariés.



**FIGURE 9 – Analyse de la corrélation des scores CNA entre les deux sous-types tumoraux du patient MpBC9 et MpBC15.** Représentation de la corrélation entre les scores CNA du tissu tumoral chondroïde (en ordonnée) et épithéial tumoral (en abscisse) pour les patientes MpBC9 (**A**) et MpBC15 (**B**). Chaque point correspond au score d'une cytobande mineure du bras chromosomique 13q. Les points rouges indiquent les cytobandes situées sur les bras chromosomiques significatifs. La ligne noire pleine représente une corrélation parfaite entre les scores CNA (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. On retrouve ces altérations divergentes sur le profil CNA (**C**).

## 4 Discussion

Les spots de la technologie Visium ont une taille d'environ 55 $\mu\text{m}$  et contiennent en moyenne 10 cellules. Bien que cette technologie permette d'allier l'information du séquençage transcriptomique à l'organisation spatiale du tissu, ce qui n'est pas possible par analyse *bulk* ou *single-cell*, cette pluralité des cellules par spot comporte certaines limitations. Cela peut en effet diluer le signal tumoral si les spots considérés comme tumoraux contiennent aussi des cellules normales. Dans notre projet, où un échantillon contient plusieurs sous-types tumoraux, cela peut mener à des « contaminations » locales très difficiles à éliminer, et à un bruit de fond pénalisant la précision des analyses transcriptomiques. Cela complique l'attribution précise d'un signal à un sous-type tumoral spécifique, et se traduit donc par un positionnement parfois diffus des spots dans l'espace vectoriel de la PCA et les projections UMAP. Il est par exemple difficile de définir une frontière exacte entre les différents sous-types mésenchymateux, fusiformes ou chondroïdes, produisant souvent de la matrice extracellulaire. Le signal transcriptomique issu de cette matrice peut artificiellement accroître les similitudes entre ces sous-types et les différents clusters se superposent entre eux.

Une solution serait d'utiliser une technique de déconvolution, permettant de connaître la composition cellulaire d'un spot et ainsi de corriger le signal pour déterminer les transcrits provenant des cellules d'intérêt uniquement. Des travaux précédents au sein de l'équipe ont déjà montré que des outils tels que RCTD (Robust Cell Type Decomposition) [20] sont assez puissants pour la déconvolution de spots Visium. Cependant, l'utilisation d'une déconvolution nécessite de comparer le profil transcriptomique de chaque spot aux profils transcriptomiques de référence caractéristiques de chaque type cellulaire. Or il n'existe actuellement aucun atlas transcriptomique de référence spécifique des différents sous-types tumoraux des MpBC.

Dans nos analyses, certains marqueurs phénotypiques identifiés montrent que nous arrivons en partie à capturer un signal biologique pertinent pour représenter certains sous-types tumoraux. Par exemple les marqueurs EPCAM et CD24 identifiés dans l'archétype correspondant aux cellules tumorales épithéliales, sont des marqueurs types des cellules épithéliales mammaires. De la même manière, le marqueur IBSP identifié pour le sous-type ostéo-sarcomatoïde correspond à une glycoprotéine de la matrice extracellulaire osseuse, tandis que le marqueur SERPINB4 identifié pour le sous-type épidermoïde correspond à un antigène connus des cancers kératinocytaires. Toutefois, on constate que pour le moment, de nombreux marqueurs identifiés par DGEA sont encore souvent soit trop peu spécifiques d'un seul sous-type, soit pas assez représentatifs de tous les patients d'un sous-type. Cela suggère donc qu'il sera nécessaire d'obtenir une meilleure précision par la suite.

Les limites de la technologie Visium peuvent également biaiser l'inférence du score CNA prédict par InferCNVPlus. L'estimation de ces scores CNA repose sur l'augmentation ou la réduction d'expression génique sur une portion de génome pour détecter des altérations du nombre de copies génomiques. Or, l'incorporation de cellules non tumorales peut « diluer » le signal tumoral, car les cellules non tumorales ont un profil génomique parfaitement diploïde. De plus, dans notre projet, tous les spots n'auront pas la même densité cellulaire ce qui dépendra du sous-type tu-

moral considéré. Par exemple, un spot contenant des cellules tumorales chondroïdes, isolées au sein de larges zones 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 copie peut s'en retrouver sous-estimé pour certains spots, faussant l'analyse. Enfin, cette technologie basée sur l'ARN est inadaptée à la segmentation précise et la quantification exact du nombre des CNA somatiques du cancer. Des approches ADN pourraient permettre de mieux caractériser l'évolution génétique entre les compartiments différents des MpBC mixtes.

Nous avons pu observer que les profils CNA des différents compartiments tumoraux appariés étaient très similaires, suggérant fortement une origine commune. Nos analyses ont tout de même permis d'identifier des altérations divergentes entre compartiments tumoraux appariés affectant des bras chromosomiques entiers. Cela suggère l'existence d'une relation clonale dans ces quelques cas, où l'un des 2 sous-types a pour origine une cellule de l'autre. Parmi tous les CNA divergents identifiées, la véritable question est de savoir si ces altérations génomiques sont des passagères, ou des drivers de la transdifférenciation dans les MpBC. Les mutations passagères correspondent à des altérations sans réel rôle fonctionnel, mais dont la fréquence élevée permet de suivre l'évolution graduelle du génome. En revanche, un CNA driver correspond à un changement du nombre de copie de gène spécifique contribuant activement à une progression de la tumeur (ici, la transdifférenciation). A l'avenir, une plus grande cohorte pourrait permettre d'identifier des CNA récurrents dans certaines transdifférenciations, et donc plus à même d'être des drivers.

Ce travail s'inscrit dans un cadre plus large d'un consortium de recherche national sur le carcinome mammaire métaplasique, comprenant le CLB à Lyon, l'institut Bergonié à Bordeaux, le Centre de Recherche en Cancérologie et Immunologie à Nantes et l'Institut Curie à Paris. Il regroupe de nombreux chercheurs et praticiens cliniques, et centralise la collecte des données cliniques et omiques des patientes atteintes de MpBC 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 [21]. Cela représente une opportunité d'enrichir notre dataset avec de nouveaux échantillons, et donc d'augmenter la représentation de chaque sous-type et la puissance statistique dans nos analyses. Cela permettra ainsi de pallier à certains manques de cette étude préliminaire.

Un des objectifs de ce projet 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 le nombre important de 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, l'utilisation de logiciel comme SCENIC [22] pourrait aider à réduire le nombre de piste d'intérêt, tout en analysant mieux les gènes régulateurs

pertinents mais qui pourraient être faiblement exprimés. Ainsi nous pourrions expliquer les profils d'expressions et de transdifférenciation observés dans chaque sous-type, selon les facteurs de transcription exprimés et les potentielles voies de signalisation dérégulées dans chaque sous-type de MpBC.

Afin de corriger la résolution insuffisante des spots Visium, 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. Pour réaliser cela, nous prévoyons de compléter les données spatiales avec des données de scRNAseq, permettant l'analyse ARN des cellules uniques pour chaque échantillon. Plus précisément cela est envisagé par single-nuclei RNAseq (séquençage des noyaux uniques), plus adapté aux échantillons fixés en paraffine (FFPE). Ces données transcriptomiques générées à l'échelle de la cellule unique permettront de créer un profil transcriptomique spécifique pour chaque sous-type cellulaire, auquel nous pourrons nous rapporter lors de l'étape de déconvolution des spots. Cela permettra à la fois de mieux évaluer la composante micro-environnementale des différents compartiments, et des éventuelles différences qui les caractérisent, mais aussi d'identifier des marqueurs sur des données « pures » au niveau de chaque type cellulaire, et donc plus spécifiques.

Dans le but d'explorer en profondeur les dynamiques évolutives des échantillons biphasiques MpBC, des microdissections par capture laser vont être effectuées 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, tels que les CNA mais aussi désormais les mutations ponctuelles, seront analysés par séquençage d'exome entier. A l'aide de ces analyses basées sur de l'ADN plus adapté que l'ARN, on pourra ainsi valider plus précisément les CNA identifiés avec mon analyse préliminaire, mais aussi de suivre plus précisément les probables relations clonales entre compartiment appariés.

Les microdissections de chaque compartiment tumoral pourront également permettre une analyse épigénétique de chaque sous-type. Nous pourrons révéler les gènes suppresseurs de tumeurs hyperméthylés et inactivés par silencing, ou au contraire des oncogènes hypométhylés favorisant la prolifération et la survie, et l'instabilité génomique des cellules tumorales. 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 [23].

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 niches biologiques au sein du tissu tumoral, pouvant réguler la trajectoire évolutive des populations tumorales [24]. Grâce à la déconvolution des spots Visium, nous pourrons 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 transdifférenciées et les cellules du micro-environnement. La spatialisation du signal dans les données de transcriptomique spatiale permettra d'identifier d'éventuels patterns d'expression ségrégée, pour déterminer si ces ligands, récepteurs et les voies qu'ils régulent présentent une expression différente entre l'environnement immédiat, la bordure et le cœur des compartiments tumoraux.

Finalement, chacun des biomarqueurs identifiés comme spécifiques de chaque sous-type fera l'objet d'une validation Immunohistochimique (IHC) dans la cohorte MpBC indépendante de l'Institut Curie, afin d'estimer leur pertinence pour le diagnostic moléculaire. Enfin, les bio-banques d'organoides et xénogreffes dérivées de patientes MpBC par les collaborateurs de l'équipe permettront d'étudier le rôle des biomarqueurs et des voies identifiées, via des perturbations génétiques et pharmacologiques.

## 5 Conclusions

Les objectifs de ce stage étaient, dans un premier 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'altérations du nombre de copie, d'identifier des altérations génétiques qui pourraient être à l'origine des transdifférenciations entre les compartiments tumoraux.

A travers l'analyse de 16 premiers échantillons MpBC, j'ai pu identifier des marqueurs phénotypiques prometteurs pour certains sous-types tumoraux. Toutefois des analyses plus poussées permettront de valider leur pertinence, et d'identifier des marqueurs spécifiques pour tous les sous-types. Mes analyses suggèrent de plus que les 2 compartiments de chaque tumeur analysée ont la même cellule d'origine. J'ai pu identifier de rares altérations suggérant une évolution clonale entre les compartiments, mais cela ne concerne qu'une partie des échantillons, suggérant que les CNAs ne sont en général pas la cause des transdifférenciations observées.

Bien que certaines limitations, liées à la résolution spatiale des spots de la technologie Visium, limitent encore la portée de mes analyses, ces premiers résultats constituent tout de même une base solide et encourageante pour la suite du projet, nous permettant d'orienter nos prochaines recherches.

Ce travail m'a également permis de consolider grandement mes connaissances dans les analyses de transcriptomique spatiale et des outils associés, avec à la fois des applications phénotypiques et génomiques.

## 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] 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.
- [3] 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.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] 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>, 2024. Consulté le 18 mai 2025.
- [8] Illumina. Unique molecular identifiers (umis). <https://www.illumina.com/techniques/sequencing/ngs-library-prep/multiplexing/unique-molecular-identifiers.html>, 2023. Consulté le 18 mai 2025.
- [9] 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.
- [10] 10x Genomics. Space ranger software (version 2.0.0). <https://www.10xgenomics.com/products/spatial-gene-expression>, 2023. Consulté le 18 mai 2025.
- [11] 10x Genomics. Loupe browser software (version 8.1.2). <https://www.10xgenomics.com/products/loupe-browser>, 2024. Consulté le 18 mai 2025.
- [12] 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 Papalexí, 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.
- [13] 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.
  - [14] Charlene Zhang. infercnvplus : Enhanced 'infercnv' package. <https://github.com/CharleneZ95/infercnvPlus>, 2020. Consulté le 18 mai 2025.
  - [15] 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.
  - [16] 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.
  - [17] David M. Suter, Nacho Molina, Damien Gatfield, Kim Schneider, Ueli Schibler, and Felix Naef. Mammalian genes are transcribed with widely different bursting kinetics. *Science*, 332(6028) :472–474, 2011.
  - [18] 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.
  - [19] Leland McInnes, John Healy, and James Melville. Umap : Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv :1802.03426*, 2018.
  - [20] D. M. Cable, E. Murray, L. S. Zou, J. Goeva, X. Macosko, E. Irizarry, P. A. Kharchenko, and E. Z. Macosko. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nature Biotechnology*, 40 :517–526, 2022.
  - [21] 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.
  - [22] S. Aibar, C. González-Blas, T. Moerman, V. Imrichova, G. Hulselmans, F. Rambow, J.-C. Marine, P. Geurts, J. Aerts, J. van den Oord, Z. Kalender Atak, and S. Aerts. Scenic : single-cell regulatory network inference and clustering. *Nature Methods*, 14 :1083–1086, 2017.
  - [23] K. Grønbaek, C. Hother, and P.A. Jones. Epigenetic changes in cancer. *APMIS*, 115(10) :1039–1059, October 2007.
  - [24] Fondation ARC pour la recherche sur le cancer. Les caf, cellules incontournables de l'environnement tumoral. <https://www.fondation-arc.org/actualites/2020/les-caf-cellules-incontournables-de-l-environnement-tumoral>, 2020. Consulté le 30 mai 2025.