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

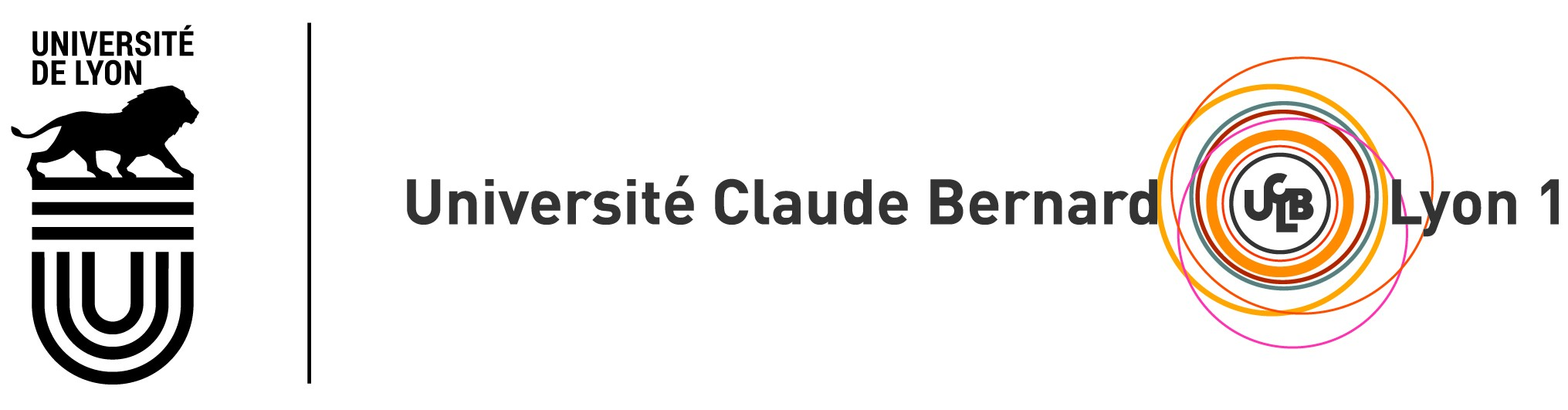
2024-2025

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**

**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 pairés mais de sous-type différent chez certaines patientes. Ces résultats contribueront à des analyses approfondies qui seront réalisées prochainement, pour déterminer les déterminants génétique et non-génétiques de la transdifférenciation dans 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**

[**Liste des abréviations** **4**](#_Toc31670)

[**Liste des logiciels utilisés** **5**](#_Toc31671)

[**1 Introduction** **6**](#_Toc31672)

[1.1 Contexte et état de l’art 6](#_Toc31673)

[1.2 Présentation des MpBC 6](#_Toc31674)

[1.3 Problématique(s) 7](#_Toc31675)

[1.4 Pertinence de la transcriptomique spatiale 8](#_Toc31676)

[**2 Matériels et méthodes** **9**](#_Toc31677)

[2.1 Echantillons MpBC 9](#_Toc31678)

[2.1.1 Cohorte : CLB 9](#_Toc31679)

[2.1.2 FFPE fixation et H&E coloration 9](#_Toc31680)

[2.1.3 Séquençage Visium & alignement 9](#_Toc31681)

[2.2 Annotation phénotypique des spots 10](#_Toc31682)

[2.3 Données scRNA-seq 10](#_Toc31683)

[2.3.1 Contrôle qualité et filtrage des spots 10](#_Toc31684)

[2.3.2 Normalisation et Scaling 10](#_Toc31685)

[2.4 Analyse des marqueurs phénotypiques 11](#_Toc31686)

[2.4.1 Harmony 11](#_Toc31687)

[2.4.2 Seurat 11](#_Toc31688)

[2.5 Analyse des altérations génétiques 12](#_Toc31689)

[2.5.1 InferCNVPlus 12](#_Toc31690)

[2.5.2 Définition des altérations génomiques divergentes 14](#_Toc31691)

[2.6 Software et packages 14](#_Toc31692)

[2.6.1 RStudio et language R 14](#_Toc31693)

[**3 Résultats** **15**](#_Toc31694)

[3.1 Analyse des marqueurs phénotypiques 15](#_Toc31695)

[3.1.1 Réduction de dimensionalité 15](#_Toc31696)

[3.1.2 Clustering et enrichissement archétypal 15](#_Toc31697)

[3.1.3 Identification des gènes marqueurs 17](#_Toc31698)

[3.2 Analyse des altérations génotypiques divergentes 19](#_Toc31699)

[3.2.1 Profils génomiques biphasiques 19](#_Toc31700)

[3.2.2 Identification des altérations divergentes entre compartiments pairés 20](#_Toc31701)

[3.2.3 Visualisation des différences de score CNA entre compartiment 21](#_Toc31702)

[**4 Discussion** **25**](#_Toc31703)

[4.1 Limites de la technologie Visium (sous-titres à supprimer) 25](#_Toc31704)

[4.1.1 Risque d’hétérogénéité cellulaire intra-spot 25](#_Toc31705)

[4.2 Perspectives du projet (sous-titres à supprimer) 28](#_Toc31706)

[4.2.1 Consortium MAESTRO : échantillons additionnels 28](#_Toc31707)

[4.2.2 Analyse de réseaux / Pathways déruglés plutôt que des gènes spécifiques 29](#_Toc31708)

[4.2.3 Analyse approfondie via snRNA-seq 29](#_Toc31709)

[4.2.4 Microdissection des compartiments 30](#_Toc31710)

[4.2.5 Déterminants non génétiques 31](#_Toc31711)

[4.2.6 Validation IHC & Analyses fonctionnelles 31](#_Toc31712)

[**5 Conclusions** **32**](#_Toc31713)

**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

1. Représentation des scores CNA obtenus avec InferCNVPlus, avant (A) et après

normalisation (B), en fonction des cytobandes mineures pour chaque chromosome

du génome . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 21

1. Profil de significativité des altérations chromosomiques pour le patient MpBC9 . 22
2. Carte de chaleur des altérations du nombre de copies pour le patient MpBC9 . . 23
3. Analyse de la corrélation des scores CNA entre les deux sous-types tumoraux du

patient MpBC9 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 24

1. 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 caRcinOma

**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)

# Introduction

## 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 [5] [6]. Enfin, ces TNBC sont caractérisés par une grande variété de sous-types, avec des phénotypes très hétérogènes [7]. Parmi eux, nous retrouvons les carcinomes du sein métaplasiques (Metaplastic Breast Cancer, MpBC) [8], 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.

## 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 un(e) pathologiste via la présence de cellules tumorales de type non-épithélial. 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 » [9]. 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 épithélial 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.

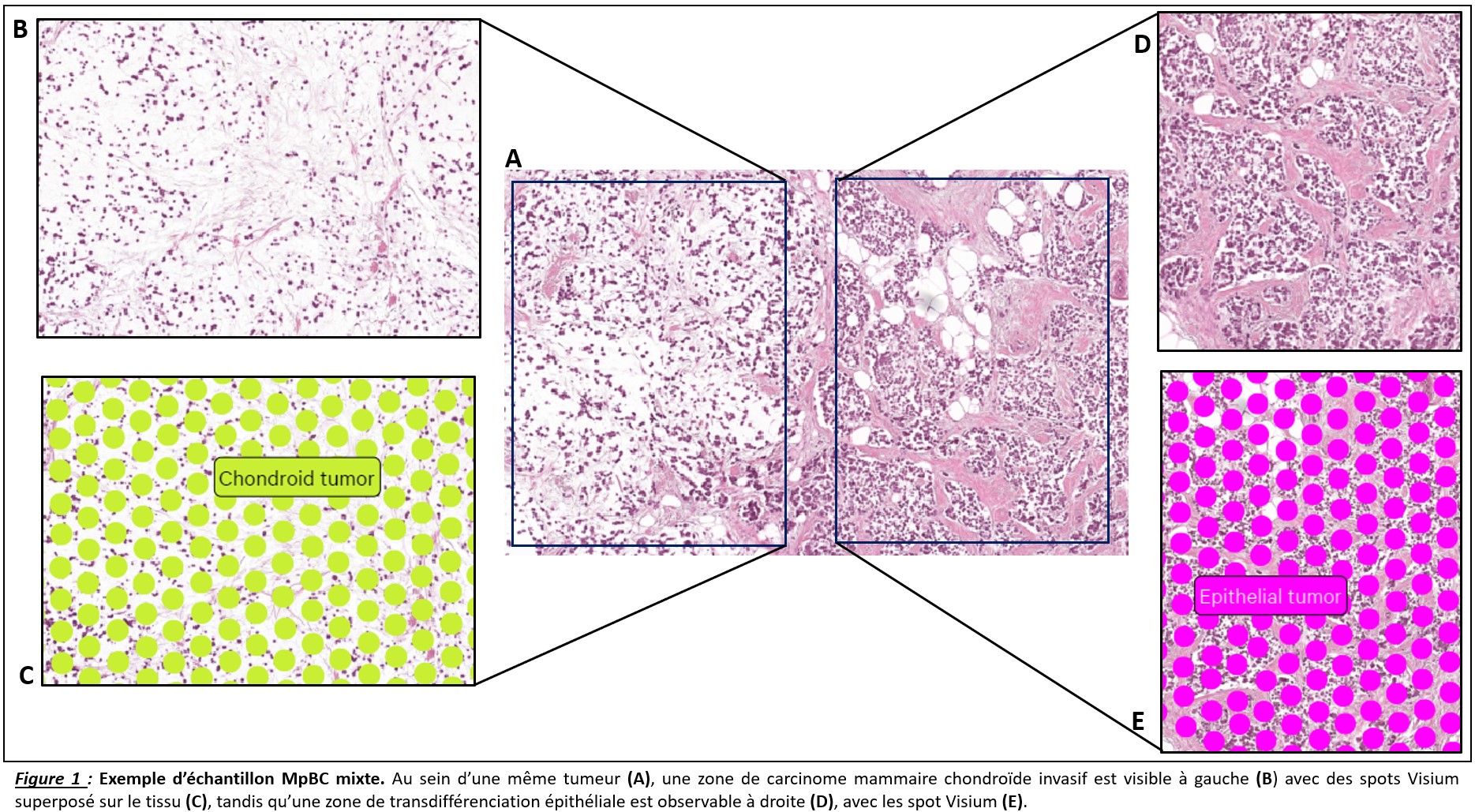




Figure 1 – Exemple d’échantillon MpBC mixte

## 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, 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.

## 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 technologie seule, tel que le bulk-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 thérapies. Enfin, cela permettra des analyses génomiques à faible résolution, mais spécifique à chaque compartiment.

# Matériels et méthodes

## Echantillons MpBC

### 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)

### 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 (Hématoxiline et Eosine). Le reste de la tumeur a été conservé à température ambiante (20°C) jusqu’à réutilisation.

### 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].

## 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 pairée 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 regroupe 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.

## Données scRNA-seq

### 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.

### 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 à 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 discrétion 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 appliqué 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.

## Analyse des marqueurs phénotypiques

### 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.

### Seurat

**UMAP** La détermination des paramètres pour réaliser la projection UMAP (Uniform Manifold Approximation and Projection), afin de réduire la dimensionalité 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 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 regardé 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 soustype 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.

## Analyse des altérations génétiques

### 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 transdifférencié 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 | Épithélial | Mésenchymateux |
| MpBC2 | Épithélial | Fusiforme (Spindle-like) |
| MpBC3 | Malpighien | Fusiforme (Spindle-like) |
| MpBC4 | Ostéosarcomatoïde | Fusiforme (Spindle-like) |
| MpBC5 | Épithélial | Fusiforme (Spindle-like) |
| MpBC8 | Épithélial | Mésenchymateux |
| MpBC9 | Épithélial | Chondroïde |
| MpBC10 | Épithélial | Fusiforme (Spindle-like) |
| MpBC11 | Épithélial | Fusiforme (Spindle-like) |
| MpBC13 | Épithélial | Chondroïde |
| MpBC14 | Épithélial | Fusiforme (Spindle-like) |
| MpBC15 | Épithélial | Chondroïde |
| MpBC16 | Épithélial | 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 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 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 (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 épithéliale). Enfin, par souci de cohérence pour la visualisation, tous les scores CNA ont été ramené à une échelle entre -1 et 1.

### 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 appliqué 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 flod-change, a voir si je peux faire le VolvanoPlot avec la figure 10 et appliquer un seuil)

## Software et packages

### 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].

# Résultats

## Analyse des marqueurs phénotypiques

### 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 afin d’analyser le profil transcriptomique de chaque échantillon. 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 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. Dans notre analyse, nous avons également tenu compte d’autres effets confondants comme ceux liés au batch de séquençage (2 batches) et au numéro de la lame de séquençage utilisée (2 échantillons par lame). On constate avec la **figure 2B** que les spots colorés, ne se séparent plus selon les patientes et se superposent en une forme similaire pour chaque patiente. 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 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 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 donnée en 50 composantes, évitant les biais techniques tout en conservant les différences biologiques que l’on souhaite étudier.

### 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 cela préserve mieux les structures locales tout en maintenant la séparation du structures locales. Le clustering Multi-Louvain a permis de déterminer les clusters de spots représentatifs des différentes annotations de sous-types tumoraux.. La combinaison de ces informations a permis d’obtenir des « archétypes », soit les spots les plus représentatifs de chaque annotation de spot / type cellulaire, illustrés en **figure 3**. On constate à nouveau une première composante discriminant les sous-types tumoraux selon un axe épithélio-

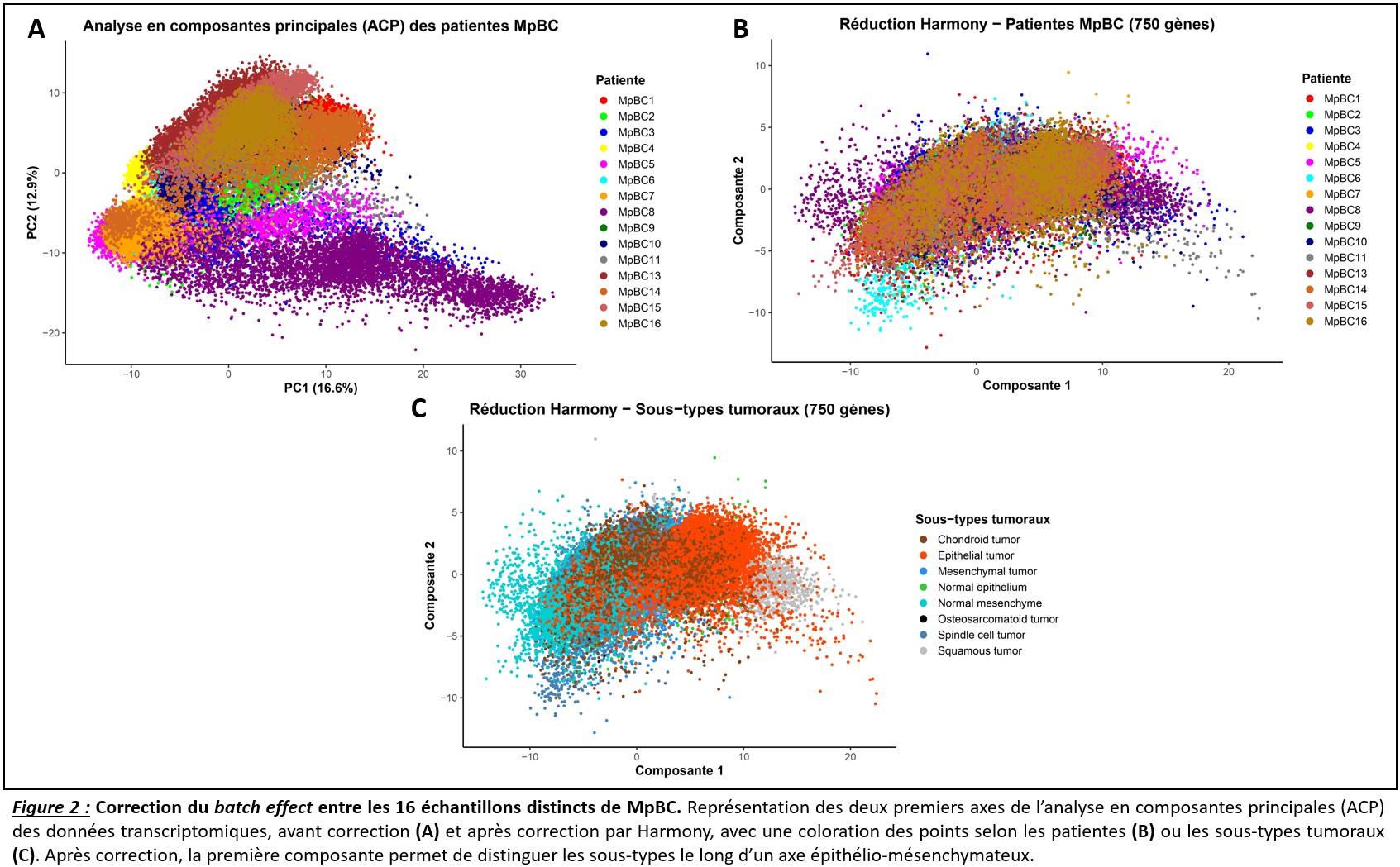


Figure 2 – Correction du batch effect entre les 16 échantillons distincts de MpBC

mésenchymateux. On retrouve le cluster des cellules tumorales malpighiennes (Squam\_tum) avec le profil le plus épithélial, prochedes cellules tumorales épithéliales (Epi\_tum). On peut également voir un cluster osteosarcomatoïde bien distinct des autres clusters et un cluster de spots annotés comme fusiformes (Spindle\_tum) qui sont à l’opposé des spotsépithéliaux. Cela reflète plutôt bien la biologie observée 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 comme mésenchymateux (Mes\_tum) se positionnent entre les 2 gros regroupements de cellule considérés comme épithéliales et fusiformes, tandis que les cellules mésenchymateuses normales se répartissent dans tous les clusters. Cela est notamment dû aux limites de la technologie Visium, qui capture par spot plusieurs cellules et donc intègrent parfois des cellules minoritaies de types cellulaires différentes 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 clusters de cellules tumorales correspondant à 5 types de transdifférenciation en plus des cellules épithéliales typiques des TNBC. Ces clusters ont par la suite été utilisés pour déterminer des marqueurs phénotypiques (gènes surexprimés) spécifiques à chacun d’eux.

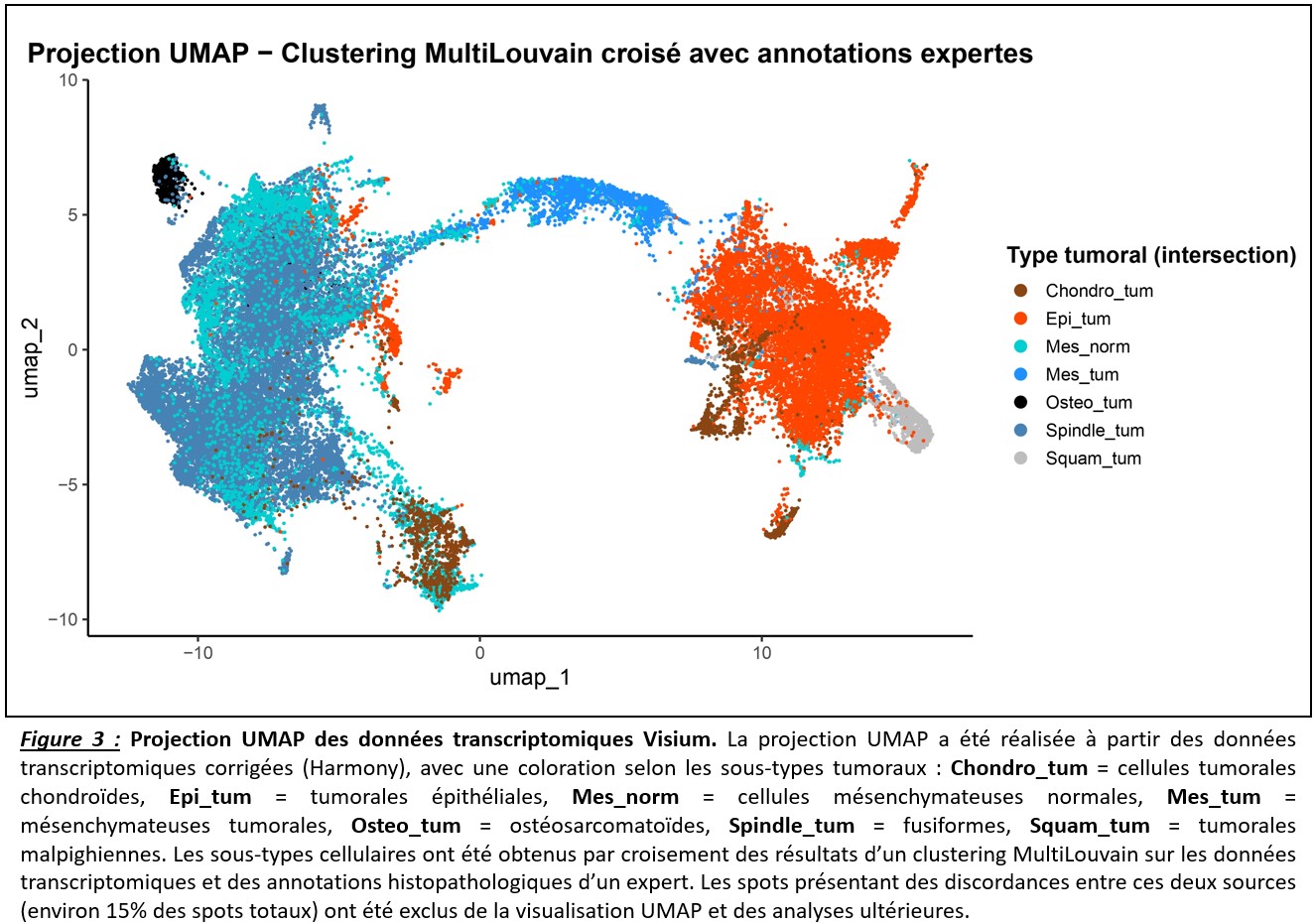


Figure 3 – Projection UMAP des données transcriptomiques Visium

### Identification des gènes marqueurs

A partir des archétypes clusters 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, nous avons notamment identifié plusieurs marqueurs phénotypiques étant spécifiquement sur-exprimés dans chacun des clusters. Les plus significatifs et biologiquement intéressants d’entre eux sont illustrés via la projection UMAP de la **figure 4**. Tout d’abord, on constate que certains archétypes sont plus difficiles que d’autre à caractériser via des gènes spécifiques. C’est le cas, par exemple, pour 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 retrouvent soit exprimés dans plusieurs autres clusters (marqueur non spécifique), soit spécifiques uniquement d’une partie des cellules d’intérêt, mais pas la totalité (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 des cellules, mais également avec une importante surexpression comparée aux autres clusters (et donc spécifique). 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 marqueurs SERPINB4 (antigène connus 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 malpighiens présentent des marqueurs surexprimés spécifiquement pour ces sous-types et sont pertinents sur le plan biologique.

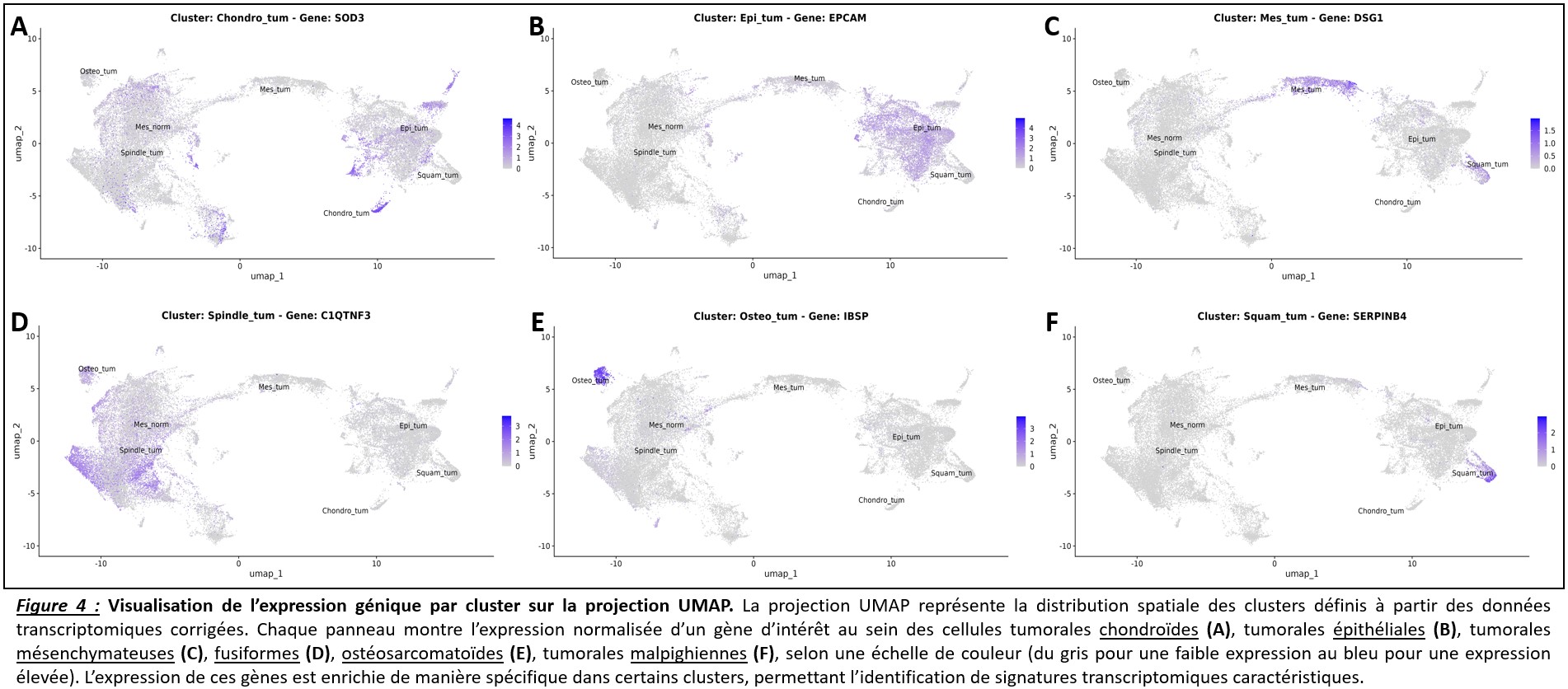


Figure 4 – Visualisation de l’expression génique par cluster sur la projection UMAP

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 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, conforte l’idée qu’il s’agit d’un candidat intéressant pour caractériser les 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 patientes présentant ce type cellulaire. De la même façon, d’autre marqueurs comme le gène SPRR1B, semble 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 présente 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 qu’au-delà d’une analyse sur la totalité des cellules de chaque archétype, il est nécessaire de réaliser une analyse plus poussée pour garantir que les marqueurs les plus significatifs ne sont pas patient-dépendants..

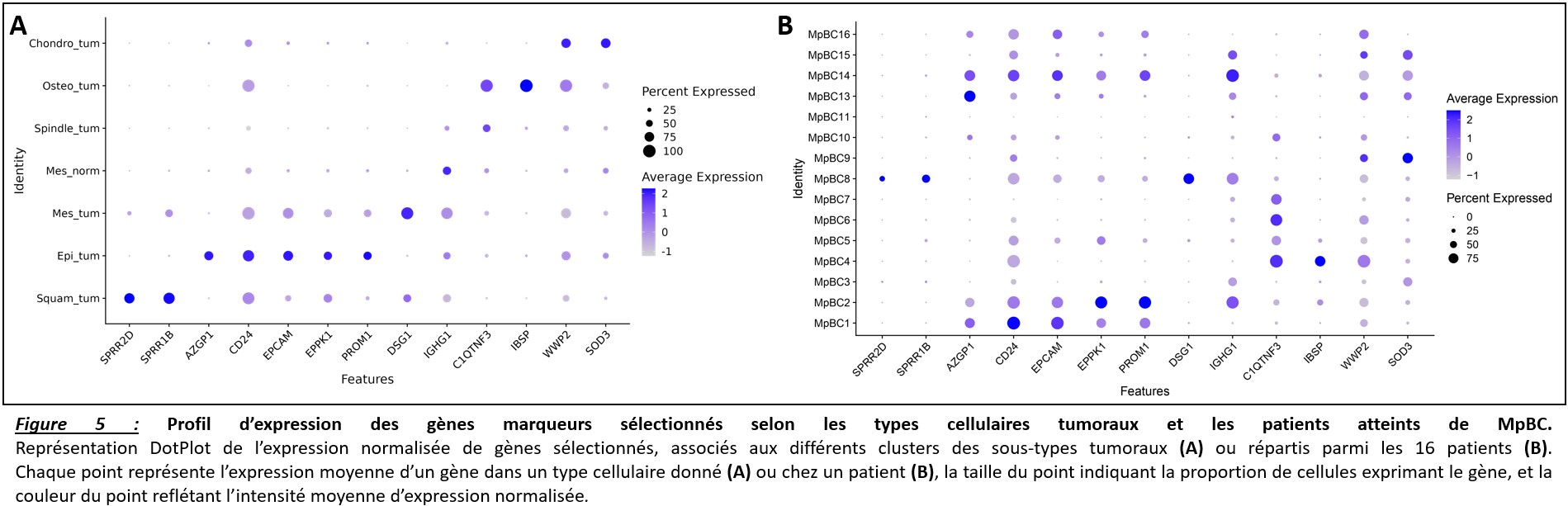


Figure 5 – Profil d’expression des gènes marqueurs sélectionnés selon les types cellulaires tumoraux et les patients atteints de MpBC

## Analyse des altérations génomiques divergentes

### 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 gène et chaque colonne correspond à un spot. Enfin, nous avons déterminé le profil médian par compartiment, en utilisant pour chaque cybtobande la médiane entre tous les spots du compartiment (Figure X). Pour obtenir le profil CNA par patiente représenté en **figure 6**, nous avons regroupé individuellement pour chaque patiente les scores CNA des gènes des matrices InferCNVPlus appartenant aux même 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 CNA par région chromosomique, cela nous a permis de baser le signal sur des régions génomiques composées de plusieurs gènes, et donc moins sensibles aux biais techniques et biologiques (variation du nombre d’UMIs entre types cellulaires, effet « burst » de l’expression ARN…).

**Normalisation par patiente** 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 différents, 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 euxà l’aide d’un facteur d’ajustement basé sur le rapport des extrêmes entre les 2 tissus (voir Méthodes). Par exemple, pour le patient MpBC9, le tissu tumoral épithélial présentait la plus forte amplification (**figure 6A**, chromosome 8), à laquelle le tissu tumoral chondroïde a été ajusté (**figure 6B**). La même correction a été appliquée pour les délétions (scores < 0). Enfin, afin de faciliter la visualisation, les scores ont été renormalisés pour couvrir une échelle de -1 à 1, comme le signal inferCNV initial. Cette normalisation a donc permis de faire coïncider les extremums 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.

### Identification des altérations divergentes entre compartiments pairés

**Test par bras chromosomique** Afin d’identifier les altérations génomiques divergentes entre les compartiments pairés 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. Pour cela nous avons défini pour bras chromosomique (p ou q) les qui le composent. Puis, grâce un test *t* (*Student’s t test*), nous avons pu tester, pour chaque patiente, si la distribution des différences entre compartiment pour chaque cytobande du bras chromosomique si elles étaient significativement différente de celles des autres bras. Nous avons ainsi obtenu une p-value pour chaque bras chromosomique.

**Résultats significatifs** Nous avons appliqué une méthode de correction stricte, correspondant à la correction de Bonferroni. Ainsi les p-values obtenues lors des tests ont été corrigées afin de fournir des valeures ajustées au grand nombre de tests efectués, avec une probabilité de commettre une erreur faible (α=0.001). En plus de ce seuil alpha, une 2e métrique plus représentative la taille de l’effet a été utilisée, à savoir la différence médiane entre les compartiments pour chaque cytobande (voir Méthodes). . En effet, en utilisant cette métrique comme filtre, nous avons pu éliminer les bras chromosomiques significatifs mais ayant une taille d’effet trop peu importante, et donc potentiellement moins intéressants.

### 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 épithélial 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 épithélial 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 complétement 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.

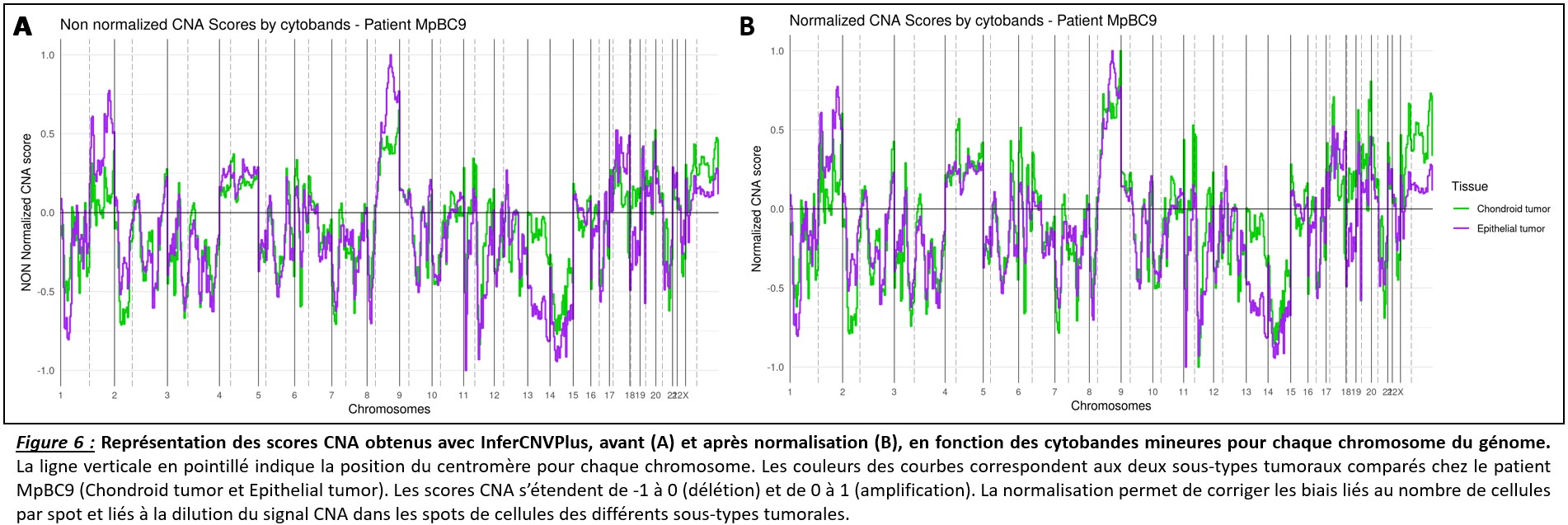


Figure 6 – Représentation des scores CNA obtenus avec InferCNVPlus, avant (A) et après normalisation (B), en fonction des cytobandes mineures pour chaque chromosome du génome

Les tests statistiques réalisé par bras chromosomiques permettent d’identifier les régions chromosomiques significativement altérées comparé aux autres régions, par 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.

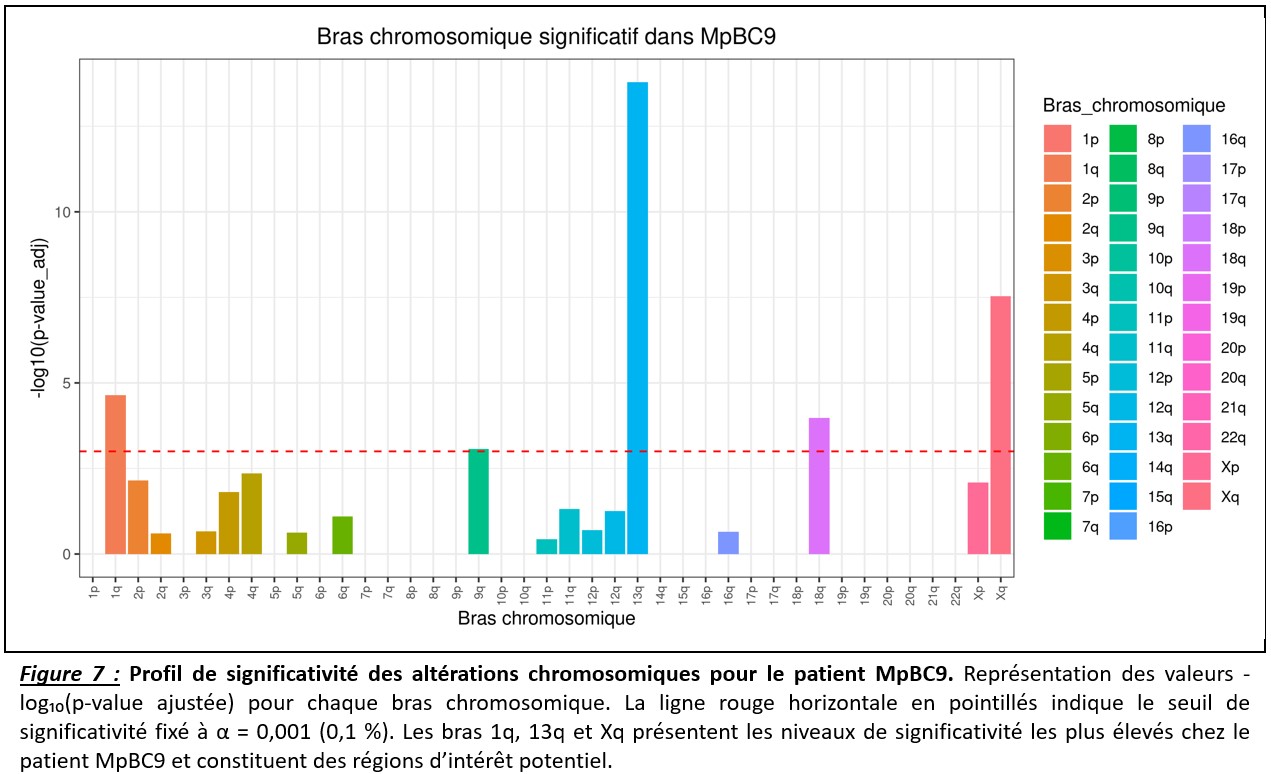


Figure 7 – Profil de significativité des altérations chromosomiques pour le patient MpBC9

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 obtenue 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 plain). 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

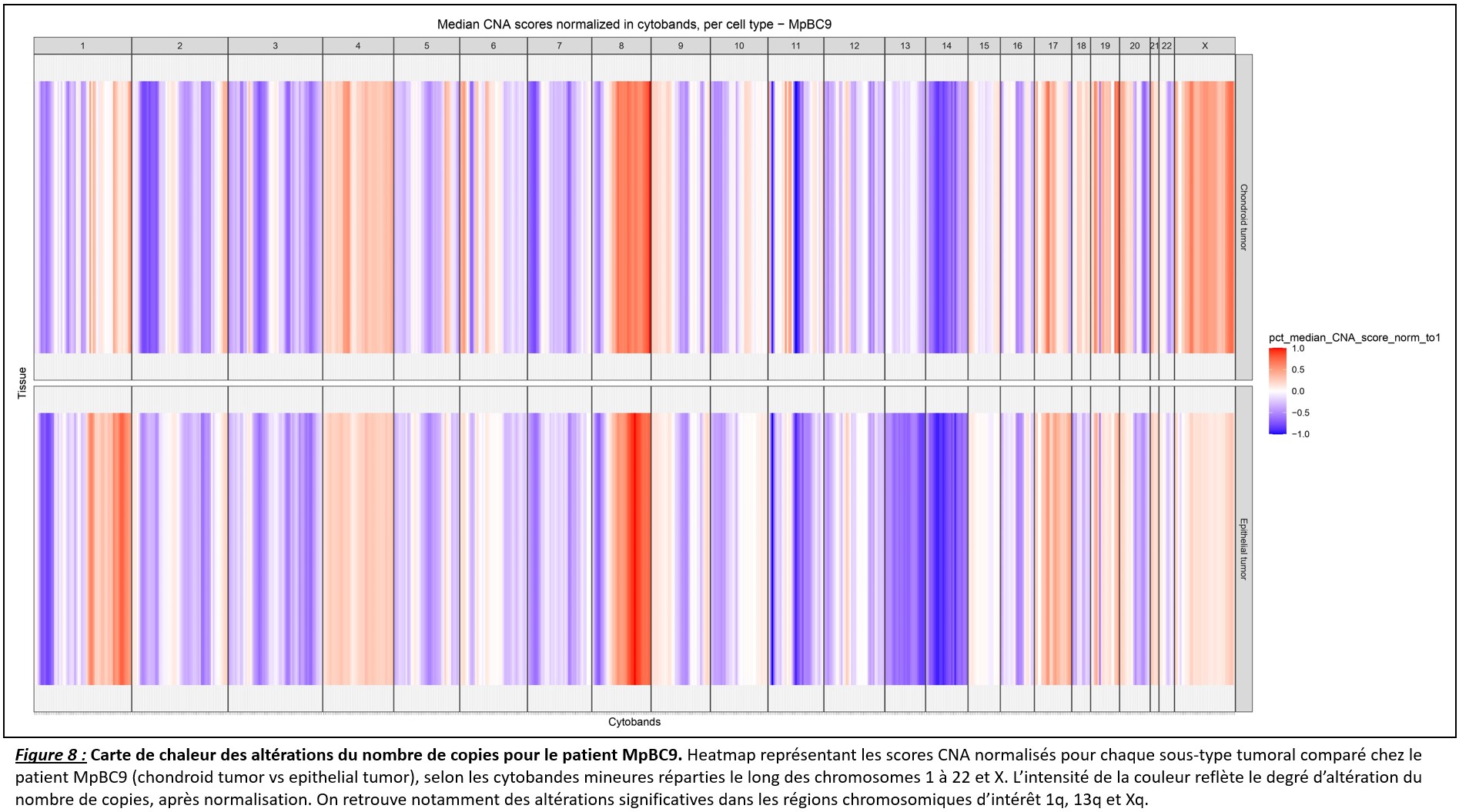


Figure 8 – Carte de chaleur des altérations du nombre de copies pour le patient MpBC9

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 appliqué 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 chromosomique 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 épithélial 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 chromosiques et comprendre les possibles causes génomiques des MpBC.

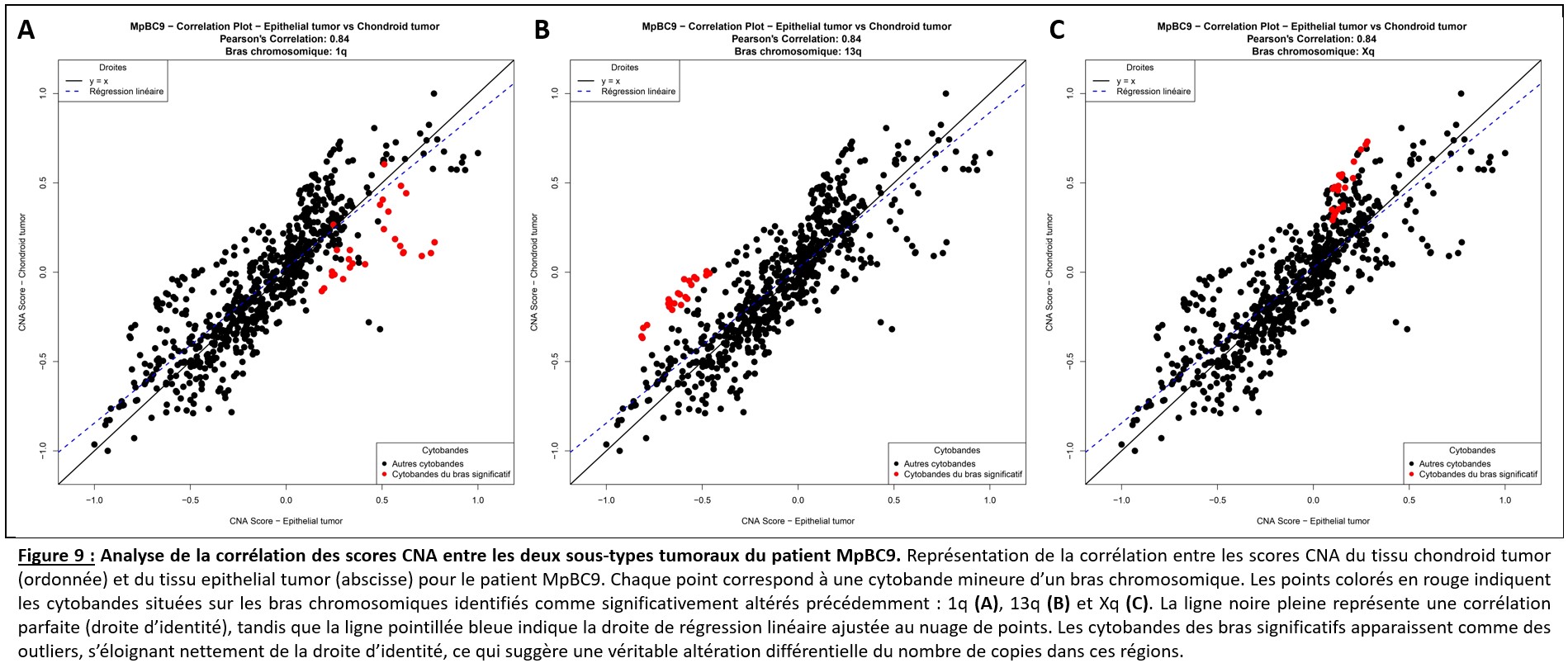


Figure 9 – Analyse de la corrélation des scores CNA entre les deux sous-types tumoraux du patient MpBC9

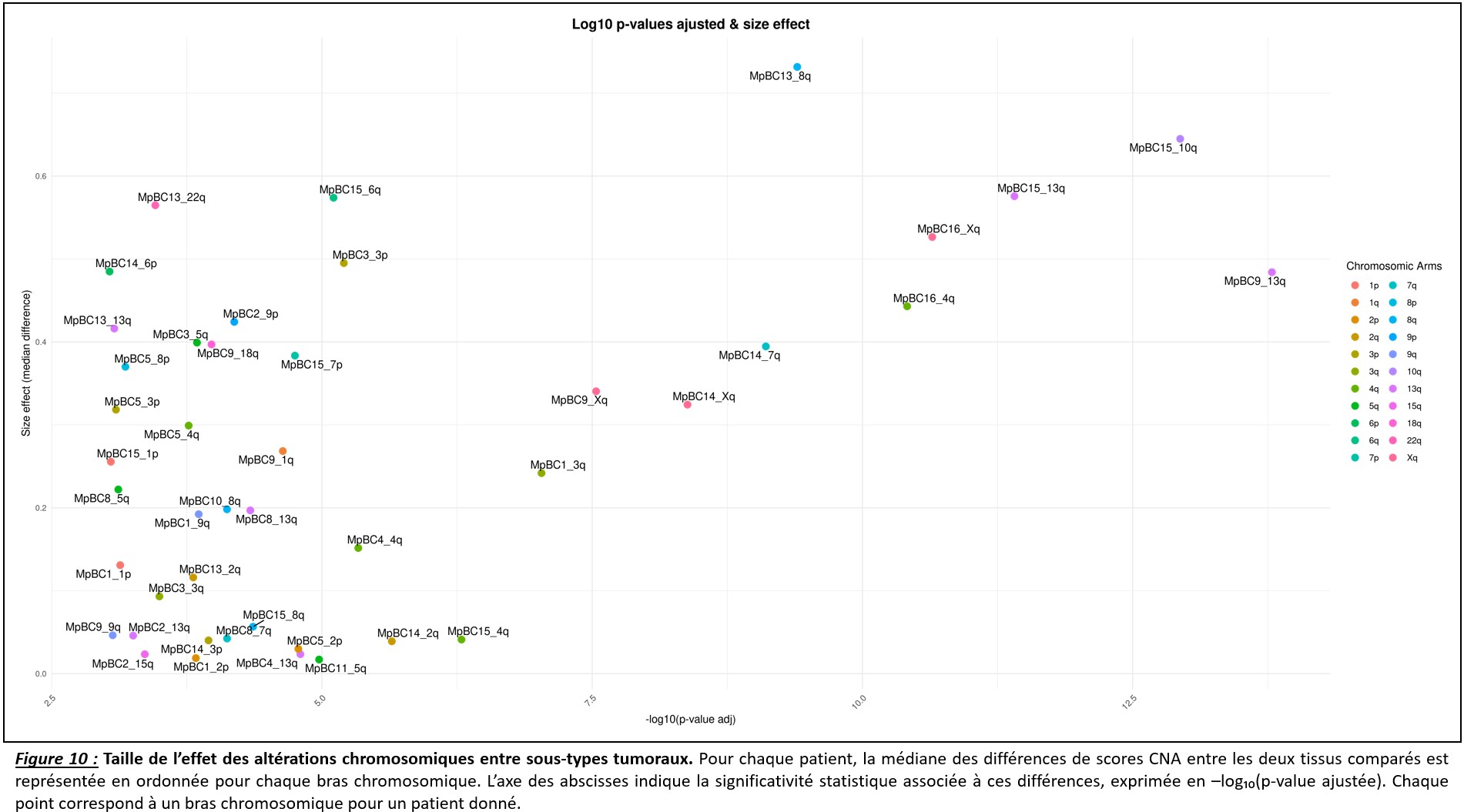


Figure 10 – Taille de l’effet des altérations chromosomiques entre sous-types tumoraux

# Discussion

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

### Risque d’hétérogénéité cellulaire intra-spot

**Séparation imparfaite des clusters (UMAP)** Lees spots de la technologie Visium ont une taille d’environ 55µm et contiennent en moyenne 10 cellules. 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 des analyses 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 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 chondroïde, 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 connaitre la composition cellulaire d’un spot et ainsi 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) 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, ersignal quepertinent pour ercertains sont des marqueurs types des cellules épithéliales mammaires Toutefois, on constate que pour le moment, de nombreux marqueurs identifiés par DGEA pour 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 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 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 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 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 affectant des bras chromosomiques entiers divergentes entre compartiments tumoraux appariés. 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 copies de gène spécifiques 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.

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

### Consortium MAESTRO : échantillons additionnels

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 [20]. 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 nos analyses. Cela permettra ainsi de pallier à certains manques de cette étude préliminaire.

### Analyse de réseaux / Pathways déruglés 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 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 (Single-Cell rEgulatory Network Inference and Clustering) pourrait aider à réduire le nombre de pistes d’intérêt, tout en analysant mieux les gènes régulateurs pertinents mais qui pourrait être faiblement exprimés. Ainsi nous pourrions expliquer les profils d’expressions et 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.

### single-nuclei RNA-sequencing

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 donnes spatiales avec des données de scRNAseq, 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 spécifique pour chaque sous-type cellulaire, que 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.

### 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 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 pairés.

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 inactivé 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. (ref)

### Déterminants non génétiques

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. 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 transdifférenciés et les cellules du micro-environnement. La spatialisation du signal dans les données TS permettra d’indentifier d’éventuels pattern d’expression ségréguée, 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.

### Validation IHC & Analyses fonctionnelles

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 biobanques d’organoïdes et xénogreffes dérivés de patientes MpBC par les collaborateurs de l’équipe permettront d’étudier le rôledes biomarqueurs et voies identifiées, via des perturbation génétiques et pharmacologiques.

# 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’altération du nombre de copie, d’identifier des altérations 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 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,](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,](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,](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.

1. BRCA France. Le cancer héréditaire du sein et de l’ovaire. [https://www.brcafrance.fr/ cancer-hereditaire-sein-et-ovaire/,](https://www.brcafrance.fr/cancer-hereditaire-sein-et-ovaire/) 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,](https://www.cancer.org/cancer/breast-cancer/about/types-of-breast-cancer/triple-negative.html) 2023. Consulté le 18 mai 2025.
3. Société canadienne du cancer. Grade et stade du cancer. [https://cancer.ca/fr/ cancer-information/what-is-cancer/stage-and-grade/grading,](https://cancer.ca/fr/cancer-information/what-is-cancer/stage-and-grade/grading) 2023. Consulté le 18 mai 2025.
4. 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.
5. 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.
6. CNRS. Est-ce qu’une division change les étapes de reprogrammation pour une cellule? [https://www.insb.cnrs.fr/fr/cnrsinfo/](https://www.insb.cnrs.fr/fr/cnrsinfo/est-ce-quune-division-change-les-etapes-de-reprogrammation-pour-une-cellule)

[est-ce-quune-division-change-les-etapes-de-reprogrammation-pour-une-cellule,](https://www.insb.cnrs.fr/fr/cnrsinfo/est-ce-quune-division-change-les-etapes-de-reprogrammation-pour-une-cellule) 2024. Consulté le 18 mai 2025.

1. 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,](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.

1. 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](https://www.10xgenomics.com/blog/spatially-resolved-transcriptomics-an-introductory-overview-of-spatial-gene-expression-profiling-methods)2023. Consulté le 18 mai 2025.
2. 10x Genomics. Space ranger software (version 2.0.0). [https://www.10xgenomics.com/ products/spatial-gene-expression,](https://www.10xgenomics.com/products/spatial-gene-expression) 2023. Consulté le 18 mai 2025.
3. 10x Genomics. Loupe browser software (version 8.1.2). [https://www.10xgenomics.com/ products/loupe-browser,](https://www.10xgenomics.com/products/loupe-browser) 2024. Consulté le 18 mai 2025.
4. 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 Papalexi, Eleni Mimitou, and Rahul Satija. Integrated analysis of multimodal single-cell data. [https://satijalab.org/seurat/,](https://satijalab.org/seurat/) 2021. Seurat R package, version 5.1.0, consulté le 18 mai 2025.
5. 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.
6. Charlene Zhang. infercnvplus : Enhanced ’infercnv’ package. [https://github.com/ CharleneZ95/infercnvPlus,](https://github.com/CharleneZ95/infercnvPlus) 2020. Consulté le 18 mai 2025.
7. Broad Institute. infercnv of the trinity ctat project. [https://github.com/ broadinstitute/infercnv,](https://github.com/broadinstitute/infercnv) 2024. Consulté le 18 mai 2025, version stable recommandée par le dépôt.
8. UCSC Genome Browser. Ucsc genome browser : Human genome assembly grch38/hg38. [https://genome.ucsc.edu/cgi-bin/hgGateway?db=hg38,](https://genome.ucsc.edu/cgi-bin/hgGateway?db=hg38) 2013. Consulté le 18 mai 2025.
9. RStudio Team. Rstudio : Integrated development environment for r (version 2024.09.0+375

"cranberry hibiscus"). [https://posit.co/download/rstudio-desktop/,](https://posit.co/download/rstudio-desktop/) 2024. Consulté le 18 mai 2025.

1. 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/,](https://www.bricbordeaux.com/en/2025/02/prix-ruban-rose-avenir-dr-monica-arnedos-team-7-bric/) 2025. Consulté le 18 mai 2025.