Page de garde

Résumé

Table des matières

Table des figures

Liste des abréviations utilisées

TNBC = Triple Negative Breast Cancr

MpBC = Metaplasic Breast Carcinoma

CNA = Copy Number Alteration

CNV = Copy Number Variant

CLB = Centre Léon Bérard

H&E = Hematoxiline & Eosine

FFPE = Formalin-Fixed, Paraffin-Embedded

CNRS = Centre Nationale de 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

Liste des logiciels et bases de données utilisés (avec numéros de versions)

RStudio

Seurat

MAST

InferCNVPlus

Loupe Browser

Space Ranger

# Introduction

## Contexte et état de l’art

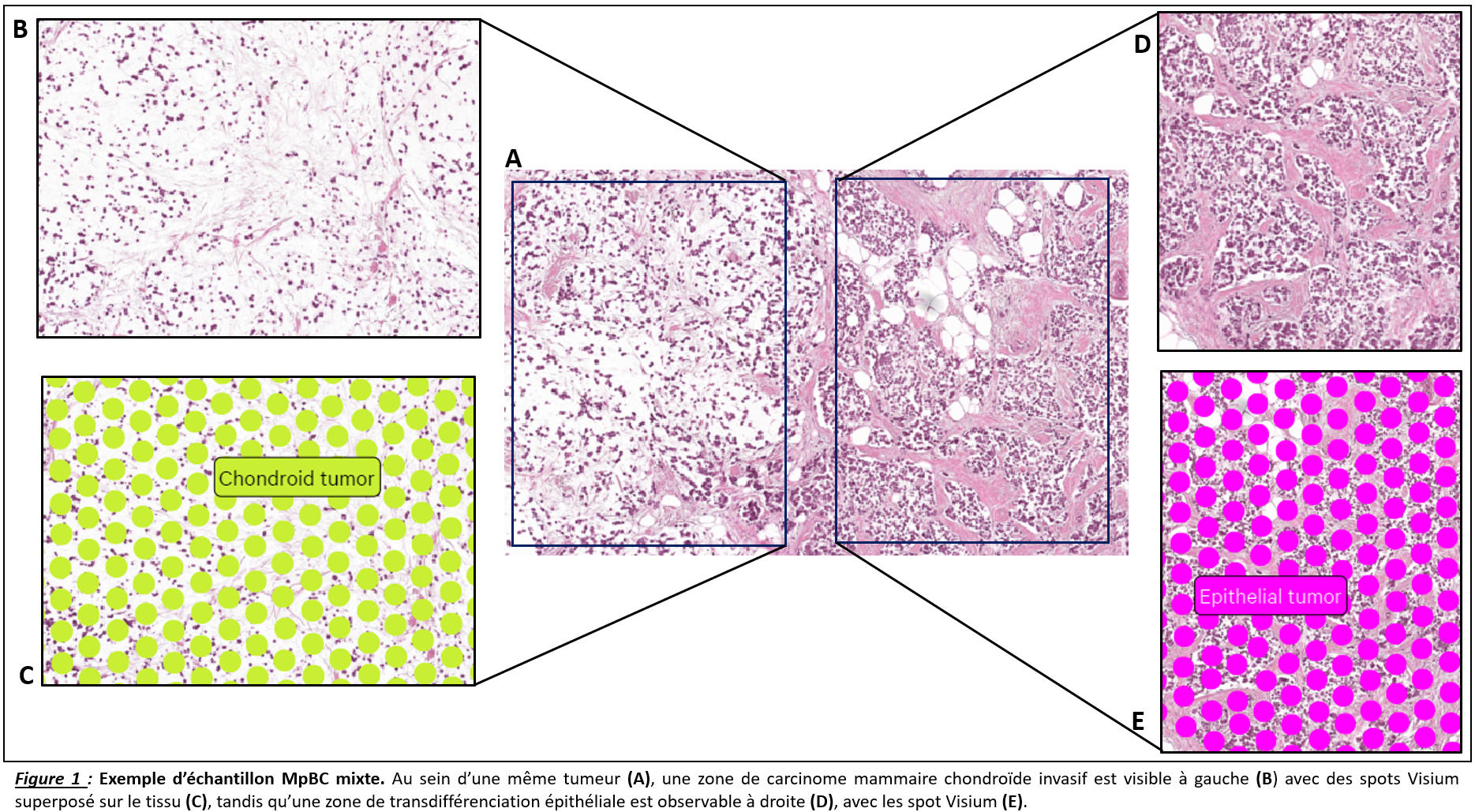
Selon l’OMS (Organisation Mondiale de la Santé) en 2022, le cancer du sein était la première cause de cancer chez les femmes dans 157 pays sur 185 [https://www.who.int/fr/news-room/fact-sheets/detail/breast-cancer], et on considère qu’environ 1 femme sur 12 en sera diagnostiqué au cours de sa vie [https://www.who.int/fr/news-room/fact-sheets/detail/breast-cancer]. Aucun facteur de risque spécifique autre que le sexe et l’âge n’est connus [https://www.who.int/fr/news-room/fact-sheets/detail/breast-cancer], faisant donc de cette pathologie un enjeu majeure de santé publique dans les années à venir. Parmi les cancers du sein, on distingue les types TNBC (Triple Negative Breast Cancer), caractérisés par l’absence de certains types de récepteurs aux œstrogènes et progestérone. Les plus connus de ces récepteurs sont les protéines HER2 (Human Epidermal growth factor Receptor 2) [https://cancer.ca/fr/cancer-information/cancer-types/breast/what-is-breast-cancer/cancerous-tumours/triple-negative-breast-cancer#:~:text=Le%20cancer%20du%20sein%20triple,ses%20propres%20options%20de%20traitement.] et BRCA1/2 (Breast Cancer 1/2) [https://www.brcafrance.fr/cancer-hereditaire-sein-et-ovaire/]. De plus, l’absence de ces récepteurs pour les patientes atteintes de TNBC rend le traitement par des thérapies ciblées extrêmement compliquées. Ainsi, les TNBC sont la plupart du temps des tumeurs agressives et de haut grade, c’est-à-dire, qu’elles présentent d’importante modifications cellulaires dysplasiques (dont le phénotype est très différent de celle des cellules normales) et disposant d’une capacité à se développer et à se propager très rapidement [https://cancer.ca/fr/cancer-information/what-is-cancer/stage-and-grade/grading]. Enfin, ces TNBC sont caractérisés par une variété impressionnante de sous-type, avec des phénotypes très hétérogènes [ref]. Parmi eux, nous retrouvons des sous-types rares, tel que les Basal-Like (Basal-Like Breast Cancer, BLBC) et les carcinomes du sein métaplasiques (Metaplastic Breast Cancer, MpBC) [ref]. Ces derniers, sont des cas rares et complexe de TNBC, encore aujourd’hui très mal compris et avec aucuns marqueurs moléculaires pour le diagnostic. Les patientes sont donc confrontées aujourd’hui à un manque important d’option thérapeutique, 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 car les cellules composant ce sous-type disposent d’une capacité remarquable à se trans-différencier. Selon le CNRS, la transdifférenciation est « la conversion d'un type cellulaire entièrement différencié en un autre type » [https://www.insb.cnrs.fr/fr/cnrsinfo/est-ce-quune-division-change-les-etapes-de-reprogrammation-pour-une-cellule]. En effet, les MpBC sont caractérisés par une importante plasticité cellulaire aboutissant à des compartiments tissulaires phénotypiquement très distinct, au sein même de la tumeur. Les échantillons MpBC présentant au moins 2 compartiments cellulaires au sein de la tumeur sont dits « mixtes », en opposition aux MpBC « pures », ne présentant pas encore de cellules cancéreuses trans-différenciées.

Actuellement, on dénombre 5 types de trans-différenciation des MpBC à partir des cellules tumorales épithéliales : Malpighienne (ou Squammeuse), Mésenchymateuse, Spindle-like (ou Fusiforme), Chondroïde et Osteosarcomatoïde. Selon le type de la trans-différenciation, les cellules tumorales peuvent présenter un phénotype, par exemple, plus épithélial que les cellules cancéreuses initiales (trans-différenciation malpighienne), ou au contraire présenter un phénotype plus stromal (trans-différenciation mésenchymateuse). Certaines cellules, lors de ce processus peuvent également prendre une forme allongée et fusiforme, en forme d’épingle (trans-différenciation Spindle-like), ou bien créer abondamment de la matrice extracellulaire autour d’elles et s’enfermer dans une logette cartilagineuse (trans-différenciation chondroïde), voire même une matrice ostéoïde ressemblant à de l’os immature (trans-différenciation ostéosarcomatoïde). Cette large diversité de différenciation des MpBC n’est, encore aujourd’hui, pas 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 aucuns traitements disponibles.

\*Rajouté une photo/figure qui montre à quoi ressemble un MpBC et des compartiments trans-différenciés



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

## Pertinence de la transcriptomique spatiale

Afin de répondre précisément à ces questions, nous avons basé nos analyses sur des données de transcriptomique spatiale généré pour les différents échantillons MpBC des patientes. Cette technologie à résolution spatiale est particulièrement intéressante dans ce projet de recherche car elle va permettre d’associer des annotations pathologiques des sous-types cellulaires, réalisé par un expert anatomopathologiste, avec des analyses multi-omiques. La nature même des MpBC faisant qu’il existe plusieurs compartiments cellulaires au sein de la tumeur, rendrait des technologies tel que le bulk-RNAseq non pertinentes dans ce contexte. Ainsi, avec une information spatiale en plus du profil transcriptomique, il sera possible d’établir des marqueurs histologiques et moléculaires spécifiques à chaque sous-type cellulaires trans-différenciés, qui pourront être utilisé par la suite pour le diagnostic et pour de potentiel nouvelles thérapies.

# Matériels et méthodes

## Echantillons MpBC

### Cohorte : CLB

Les échantillons MpBC proviennent d’une cohorte de 39 patientes diagnostiquées pour un cancer du sein et ayant un suivi régulier au Centre Léon Bérard (CLB). Chacune des patientes présentent des MpBC mixtes, ou bi-phasiques (sauf MpBC4 et 5) avec des trans-différenciations de différents types. La participation au projet de recherche s’est faite avec la non opposition des patientes et dans le cadre éthique du recueil et l’analyse de matériels biologiques imposé par l’hôpital CLB. Pour ce travail de recherche, les analyses se sont concentrées sur les 16 premiers échantillons MpBC (MpBC1 à 16).

### FFPE fixation et H&E coloration

Une fois les prélèvements cliniques réalisé pour le diagnostic, les biopsies ont suivies un protocole FFPE (Formalin-Fixed, Paraffin-Embedded) de fixation et d’inclusion en paraffine afin de conserver durablement l’échantillon et de faciliter sa réutilisation à des fins de recherche. Pour chaque patientes, 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 placé sur lame de séquençage « Visium V2 » de la société privé 10X Genomics, par paires (2 MpBC par surface de capture). Pour notre projet nous avons utilisé la technologie Visium, permettant d’apporter une information spatiale via des spots de cellules (1 spot représentant environ 10 cellules, 55µm environ), en plus des données transcriptomiques, pour chaque échantillon. 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 [https://www.illumina.com/techniques/sequencing/ngs-library-prep/multiplexing/unique-molecular-identifiers.html]. Une surface de capture peut capturer jusqu’à 18000 gènes. Pour notre projet, le séquençage s’est fait avec le kit « Visium V4 Slide - FFPE v2 » et à 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 optimales (MpBC12), nous l’avons donc exclu des analyses.

Pour chaque échantillon séquencé, les lectures ont ensuite été démultiplexées et alignées contre le génome humain (version GRCh38) via un pipeline entièrement automatisé dirigé par la plateforme bio-informatique du CLB, via le logiciel Space-Ranger (version 2.0.0).

## Annotation phénotypique des spots

Afin de disposer de l’information spatiale, nous avons annoté chacun des spots pour chaque échantillon avec les sous-types cellulaires observés sur la lame avec la coloration H&E. Pour cela nous avons combiné les informations du sous-type cellulaire fournies par 2 méthodes différentes.

Une première annotation des spots a été réalisé par le logiciel « Loupe Browser » de la société 10X Genomic, permettant de clustériser les spots présentant des profils transcriptomiques similaires, via un clustering k-means. La version de ce logiciel est Loupe Browser v8.1.2 (18 Nov 2024).

Puis dans un deuxième temps, à l’aide de ces clusters trouvés, les spots ont été individuellement vérifiés et manuellement annoté par un expert anatomopathologiste de la plateforme histologique du CLB. 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 à été réalisés. Dans un premier temps, pour chaque échantillon MpBC, la matrice de comptage 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 (problème de séquençage et cellules endommagées). Une limite supérieur correspondante au 99e percentile de l’échantillon a été appliquée pour retirer les cellules anormalement actives. Dans un second temps, les spots de cellules ont été filtré selon le pourcentage de gène mitochondriaux qu’ils contenaient, afin d’éliminer de nos données les spots de cellules contenant du tissu nécrosé ou des cellules apoptotiques, qui pourraient biaisées notre analyse. Pour cela un filtre, éliminant tous les spots contenant plus de 5% de gène mitochondriaux, a été appliqué sur tout le dataset. Ces seuils de filtrages proviennent de recommandations fournis par la société 10X Genomic pour ce genre de technologies [https://www.10xgenomics.com/analysis-guides/common-considerations-for-quality-control-filters-for-single-cell-rna-seq-data], et de conseils préconisés par la communauté scientifique utilisant ces outils (Seurat, version v5.1.0).

### 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 patientes (au préalablement filtrés des spots de mauvaises 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). Cette log-normalisation permet à la fois de compenser les potentielles différences de profondeur de séquençages entre les différents spots des différents échantillons MpBC, mais également à stabiliser la variance des comptes UMI en réduisant l’impact des spots avec des comptes UMI extrêmes.

De plus, lors de l’étape de scaling des count d’UMI, nous avons filtré le jeu de donnée afin de ne conserver que les 750 gènes les plus variables entre nos spots. Comme le recommandait la documentation du package Seurat et ses auteurs, ce seuil a été déterminer de façon empirique, car ce sont des paramètres qui sont avant tout propres et spécifiques à chaque dataset. 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.

## 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). 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 préciser 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) 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 par Harmony, avec un nombre de voisin (*n.neighbors*) de 75, une distance minimale entre les points de l’espace (*mindist*) de 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 » (*metric*) pour mesurer la distance entre les cellules sans l’espace de dimension réduit.

#### Clustering

Afin de réaliser le clustering des spots de cellules présentant les mêmes profils transcriptomiques, 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 50 premiers axes de cette 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, déterminer de façon empirique et fixé à 0.15.

#### DGEA

Une analyse des gènes différentiellement exprimés (DGEA) a été réalisé pour chaque cluster afin de déterminer les cibles moléculaires d’intérêt pour chaque sous-type tumoral. Pour cela, nous avons utilisé la fonction *FindAllMarkers()* de Seurat, avec l’option MAST (Model-based Analysis of Single cell Transcriptomics) comme test pour identifier, parmis tous les gènes, ceux étant les plus différentiellement exprimés. De plus, afin de filtrer les résultats, nous avons récupérer uniquement les gènes 4 fois plus sur-exprimés en appliquant un seuil de logFC (log Fold-Change) *(logfc.treshold = 2*). Dans notre projet, nous ne récupérons que les gènes sur-exprimés car nous cherchons des cibles moléculaires pour le diagnostic. Les gènes sous-exprimés nous intéresse donc pas ici. Enfin, un dernier filtre à 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.

## Analyse des altérations génétiques

### InferCNVPlus

#### Constitution du groupe de cellule de références

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 spot de cellule de référence à été construit à partir des spots annotés comme non tumoraux dans chaque échantillons MpBC. Comme chaque échantillons MpBC ne contenaient pas systématiquement des spots de cellules annotés comme non tumoraux, nous avons créé un groupe de référence en combinant tous les compartiments des échantillons MpBC disposant de tissu normal et 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 fournit 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 (ref)). Chaque gène présent dans la matrice de comptage initiales a été attribués à une des cytobandes mineures du génome humain. Les scores CNA ont été agrégés en faisant la médiane des scores pour chaque gène appartenant à chacune des cytobandes mineures, pour chaque patient, et chaque sous-type tumoral. Ainsi, pour chaque cytobande mineure de chaque patient, nous avions un score CNA selon les 2 types de compartiment tumoral considéré dans l’échantillon MpBC, le long de tout le génome humain. Les échantillons MpBC pures (MpBC6 et 7), ne contenant par définition qu’un seul compartiment trans-différencié, ont été exclue de l’analyse.

Voici le tableau récapitulatif des comparaisons de score CNA des tissus tumoraux par patient :

|  |  |  |
| --- | --- | --- |
| **Patients** | **Sous-type tumoral 1** | **Sous-type tumoral 2** |
| *MpBC1* | Epithélial | Mésenchymateux |
| *MpBC2* | Epithélial | Fusiforme (Spindle-like) |
| *MpBC3* | Malpighien | Fusiforme (Spindle-like) |
| *MpBC4* | Ostéosarcomatoïd | Fusiforme (Spindle-like) |
| *MpBC5* | Epithélial | Fusiforme (Spindle-like) |
| *MpBC8* | Malpighien | Mésenchymateux |
| *MpBC9* | Epithélial | Chondroïd |
| *MpBC10* | Epithélial | Fusiforme (Spindle-like) |
| *MpBC11* | Epithélial | Fusiforme (Spindle-like) |
| *MpBC13* | Epithélial | Chondroïd |
| *MpBC14* | Epithélial | Fusiforme (Spindle-like) |
| *MpBC15* | Epithélial | Chondroïd |
| *MpBC16* | Epithélial | Chondroïd |

***Table 1 :*** *Tableau récapitulatif des différents tissus (sous-type tumoral) comparés dans l’analyse des CNA par patient*

#### Normalisation

Les spots de la technologies Visium correspondent à environ 55µm de diamètre et peuvent donc contenir plusieurs cellules tumorales et non tumoral au sein d’un même spot. Par conséquent, la présence dans les spots, annotés tumoraux, de cellule non tumorale peut biaiser ces scores CNA. Afin de normaliser ces scores CNA entre les différents compartiments tumoraux pour chaque patiente, nous avons factorisé chaque scores CNA par un facteur d’échelle permettant de faire coïncider les extremums des scores de délétion (score CNA < 0) et d’amplification (score CNA > 0) entre les différents compartiments tumoraux. Par patientes, un facteur d’amplification était calculé en rapportant le score CNA maximal du tissu 1 sur le score maximal du tissu 2. De façon identique, un facteur de délétion a été calculé, en prenant cette fois-ci la valeur absolue du rapport des valeurs minimales entre tissus. Uniquement les scores CNA du tissu ne présentant pas le maximum global (tous tissus confondus) étaient multipliés par ce facteur d’échelle. Ainsi, les biais liés à la profondeur de séquençage ne perturbent plus l’analyse. Un spot contenant peu de cellule et donc une faible profondeur de séquençage (par exemple le sous-type chondroïde) est facilement comparable à des spots où la profondeur est plus conséquente (sous-type épithéliale). Enfin tous les scores CNA ont été ramené à une échelle entre -1 et 1.

### Définition des altérations génomiques divergents

#### Tests statistiques

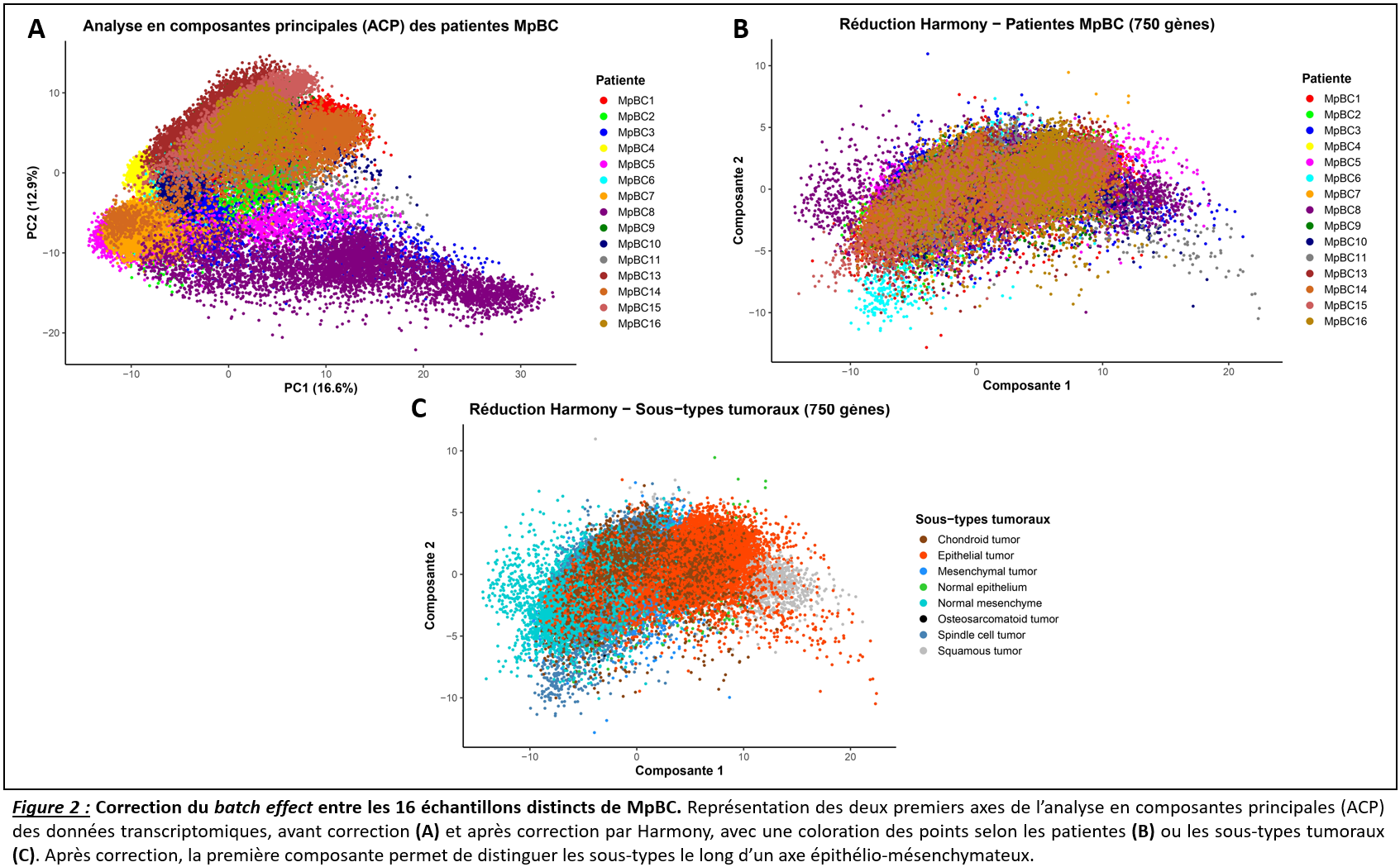
Enfin, afin de tester la significativité des altérations génomiques divergentes entre les différents sous-types tumoraux pour chaque patient, nous avons regardé si la distribution des scores CNA de toutes les cytobandes mineures pour un bras chromosomique spécifique (par exemple le bras p du chromosome 1 : 1p) était significativement différente de la distribution des scores CNA de toutes les autres cytobandes du génome, pour chaque patient individuellement. Si les distributions étaient normalement distribuées (test de shapiro supérieur à un seuil 5%), alors un test t de Student était réalisé, et le seuil α choisi était de 0.1%, sinon nous effectuionsun test non paramétrique de Wilcoxon. Une correction de Bonferronni a été appliqué sur chacune des p-value du test statistique afin de corriger les biais liés aux test multiples.

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 »).

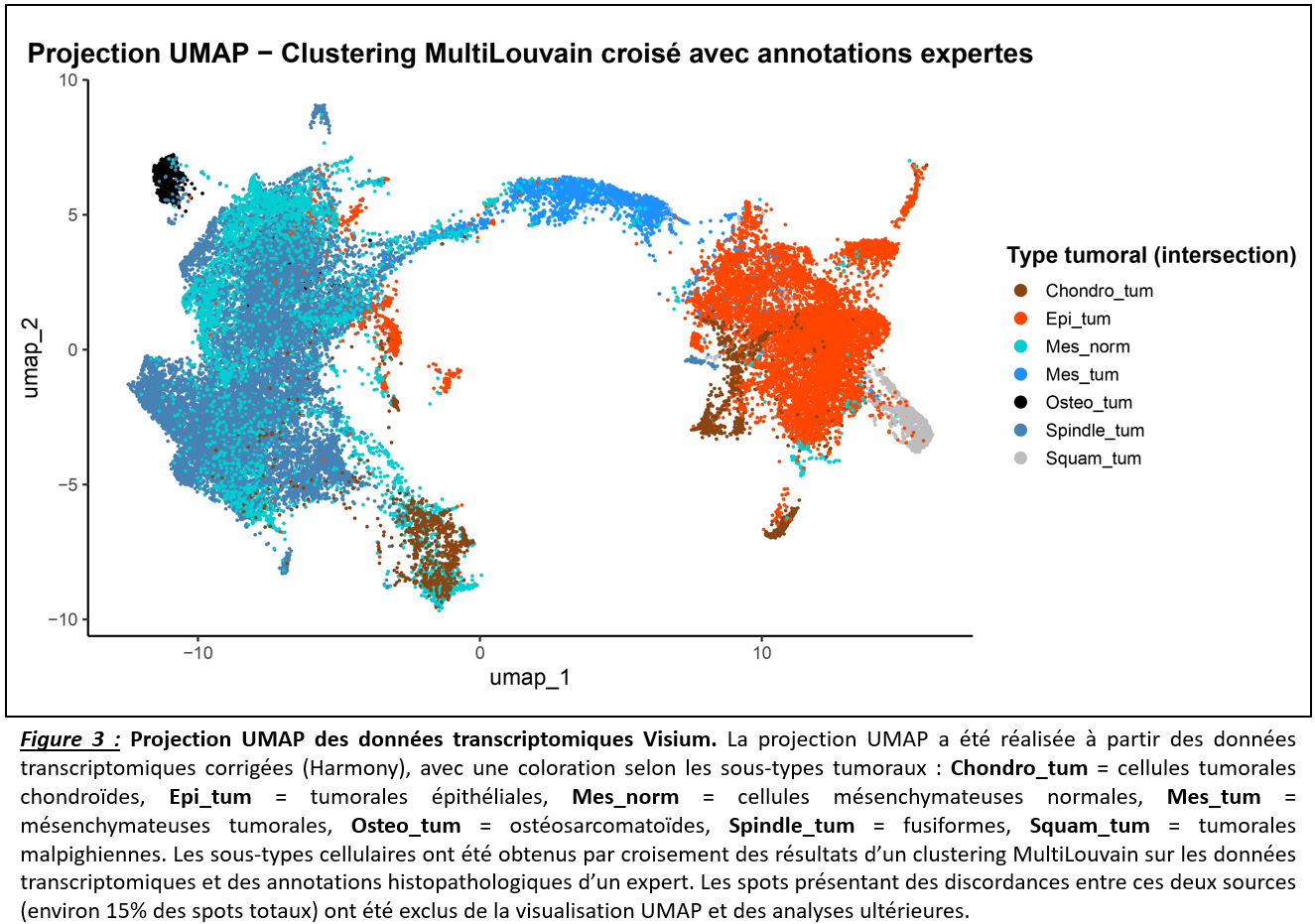
# Résultats

## Analyse des marqueurs phénotypiques

### Réduction de dimensionalité

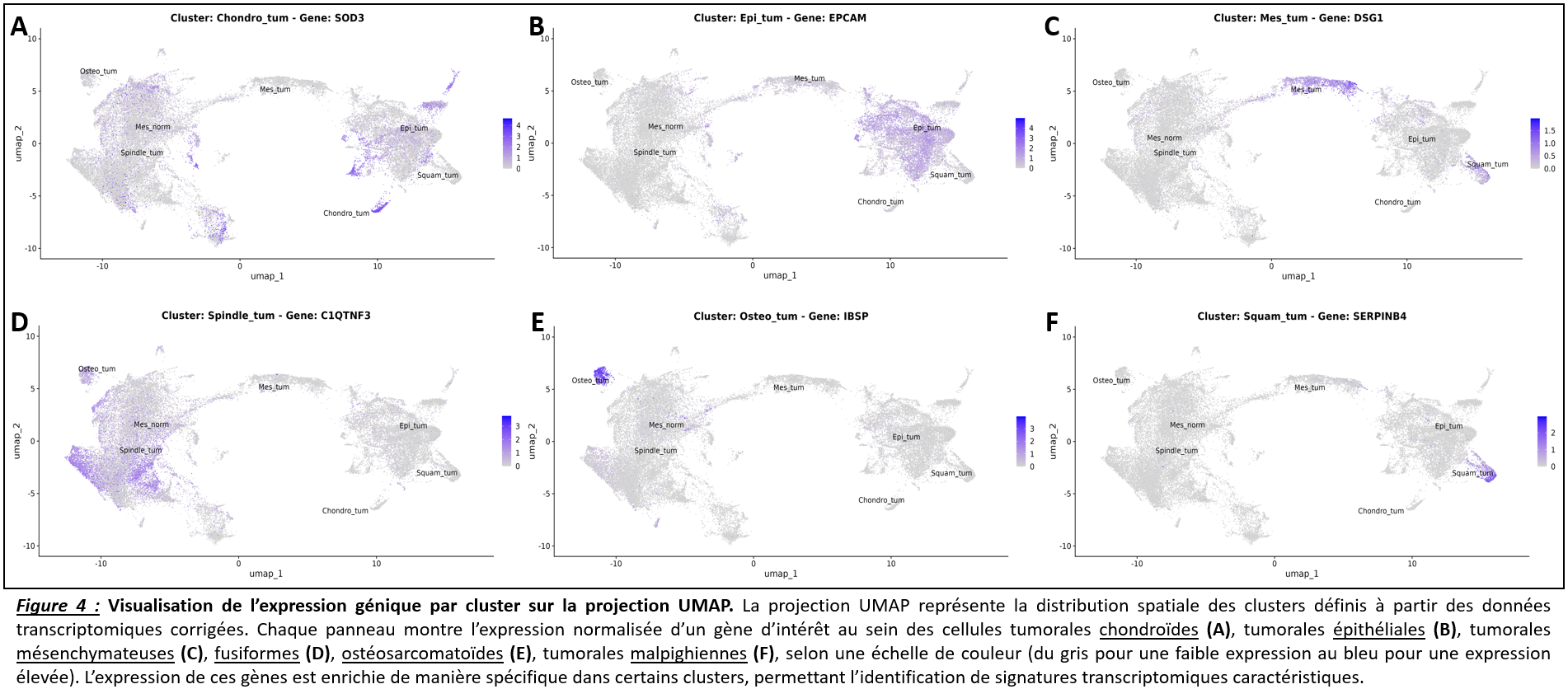
Manipuler les matrices de comptes des données transcriptomiques Visium pour 16 patients est complexe. Afin d’analyser le profil transcriptomique de chaque échantillon, une première étape de réduction de dimension linéaire est réalisée grâce à une ACP. La **figure** **2A** représente les deux premiers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

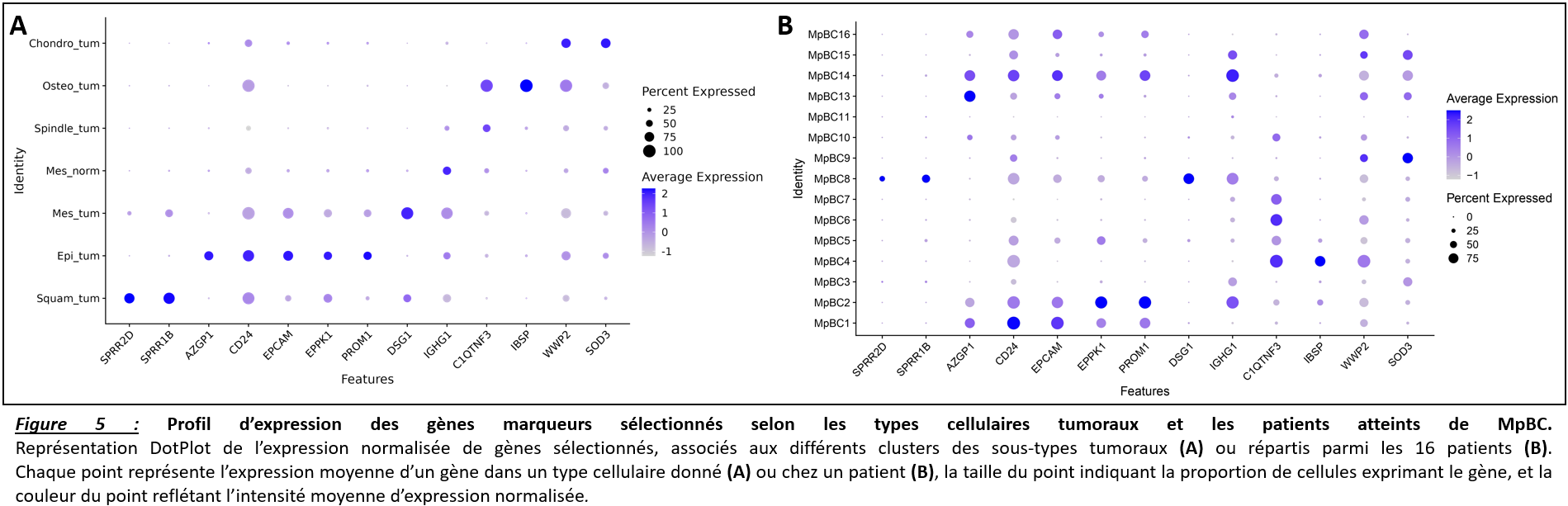
### Clustering et enrichissement archétypal

Nous avons utilisé les composantes principales corrigées des effets de lots issues de Harmony comme base de projection pour la visualisation UMAP. La UMAP est préféré à la t-SNE car elle préserve mieux les structures locales tout en maintenant la séparation du structures locales. Le clustering Multi-Louvain a permis de déterminer les clusters de spot de cellules dans l’espace vectoriel de la UMAP. Cela nous a permis de déterminer plusieurs clusters représentant les différentes annotations de sous-types tumoraux. Chacun de ces clusters ont été purifié des spots de cellules dont les annotations rentraient en contradiction avec l’annotation phénotypique de l’expert anatomopathologique. La combinaison de ces informations a permis d’obtenir la **figure 3**, représentant la projection UMAP des différents clusters de spot de cellule, correspondant à nos différents sous-types tumoraux retrouvés dans nos 16 échantillons MpBC. On constate à nouveau une première composante, discriminant les sous-types tumoraux selon un axe épithélio-mésenchymateux. On retrouve bien le cluster des cellules tumorales malpighiennes (Squam-tum) avec un profil « plus épithéliale » que les cellules tumorales épithéliales (Epi-tum). On peut également voir un cluster osteosarcomatoid bien distinct des autres clusters et un cluster de spot de cellules annotés fusiform (Spindle\_tum) qui sont, sur l’axe, du côté opposé aux spot de cellules avec un profil épithéliales. Toutefois, on constate que certains clusters restent difficiles à caractériser et se fragmentent en plusieurs sous-clusters. C’est notamment le cas pour le cluster des cellules tumorales fusiformes (Spindle\_tum) et tumorales chondroïdes (Chondro\_tum), qui sont plus difficiles à caractériser. Enfin, les spots annotés associés aux cellules tumorales mésenchymateuses se positionnent entre les 2 gros regroupements des cellules plutôt épithéliales et des cellules conjonctives, tandis que les cellules mésenchymateuses normales se répartissent et « contaminent » un peu 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 de types cellulaires différentes de l’annotations attribué pour le spot. Pour conclure, cette projection UMAP faite à partir des données transcriptomiques des 16 patients MpBC permet de caractériser 6 clusters de cellules tumorales correspondant à nos 6 types de transdifférenciation dans les MpBC. Ces clusters pourront par la suite être utilisé pour déterminer des marqueurs phénotypiques (gènes) spécifique à chacun d’eux.

### Identification des gènes marqueurs

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



Afin de comprendre plus en détail, l’expression de ces marqueurs dans ces clusters et leur pertinence pour notre problématique, nous avons regardé l’expression moyenne et le pourcentage exprimé dans les cellules des meilleurs marqueurs selon chacun des clusters identifiés (**figure 5A**) ou bien selon le patient d’origine (**figure 5B**). Nous pouvons constater sur la **figure 5A** que certains de ces marqueurs, comme le gène EPCAM, qui comme vue précédemment semble pertinent pour caractériser le cluster des cellules tumorales épithéliales, et est également surexprimé dans la majorité des patients présentant ce type de transdifférenciation dans les tumeurs. En revanche d’autre marqueurs comme le gène SPRR1B, pourrait être un marqueur pertinent pour caractériser le cluster des cellules tumorales malpighienne (**figure 5A**). Cependant, en regardant la **figure 5B**, on constate que ce marqueur n’est exprimé uniquement pour 1 seul patient (MpBC8). Or le patient MPBC3 exprime aussi ce sous-type tumoral mais on ne retrouve pas d’expression de ce marqueur pour ce patient. Cela nous montre donc que certains marqueurs identifiés, comme EPCAM, semblent être de bons candidats pour caractériser spécifiquement un des sous-types tumoraux. Toutefois certains marqueurs retrouvé significativement surexprimé dans des clusters, sont en fait patient-dépendant.

## Analyse des altérations génotypiques divergentes

### Profils génomiques biphasiques

#### Réduction par cytobande

Les résultats fournis par le package InferCNVPlus correspondent à des matrices représentant les variations du nombre de copies (CNV), où chaque ligne correspond à un gène et chaque colonne correspondent à un spot. A chaque position est attribué un score d’altération du nombre de copie par rapport à un groupe de cellule de référence. Ce score est compris entre -1 et 1 : 0 signifie qu’il n’a pas été détecté de CNA pour ce gène dans les cellules tumorales, par rapport aux groupes de référence, +/- 1 signifie que ce gène est très probablement sujet à un CNA (délétion si score négatif, amplification si score positif). Pour obtenir le profil de CNV par patient représenté en **figure 5**, nous avons regroupé, pour chaque patient individuellement, 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 score CNA, plutôt que la moyenne. Cela permet de réduire le nombre de données CNA en cytobandes mineures, tout en restant représentatif de la majorité des données. De plus, en lissant le signal CNV par région chromosomique permet d’éviter que certaines cellules ou gènes hautement amplifiés (ou délétés) ne faussent le signal.

#### Profils moyens par compartiment

Ensuite, pour avoir un profil CNV spécifique à chaque compartiment tumoral ???

#### Normalisation par patiente

Nous travaillons ici avec des spots de cellules, contrairement à ce qu’on a classiquement avec des cellules uniques. De plus il est important à annoté que chaque spot aura des profondeurs de séquençage différents, selon le sous-type tumoral considéré. En effet, nous devons donc procéder à une étape de normalisation afin de corriger les scores des spots CNA entre eux. Pour cela, à partir des médianes de scores CNA par cytobandes mineures pour chaque patient, nous avons identifié le minimum (délétion maximale) et le maximum (amplification maximale) de chacun des 2 compartiments tumoral de l’échantillon bi-phasique. Nous avons calculé un facteur d’ajustement basé sur le rapport des extrêmes entre les 2 tissus et ajusté les scores CNA du tissu « le moins extrêmes ». Par exemple, pour le patient MpBC9 de la **figure 5**, c’est le tissu tumoral épithélial qui présentait la plus forte amplification (**figure 5A**, chromosome 8), on a donc ajusté les amplifications (scores CNA > 0) du tissu appairé, le tissu tumoral chondroïde. La même transformation a été appliquée pour les délétions (scores < 0). Enfin, afin de faciliter la visualisation, les scores ont été ramenés sur une échelle de -1 à 1. Cela a donc permis, visuellement de faire coïncider les extremums des scores de délétion et d’amplification entre les différents compartiments tumoraux (**figure 5B,** chromosome 8), et, *in fine*, de limiter les biais aux différences de profondeurs de séquençage entre les différents spots annoté pour les différents sous-types tumoraux.

### Identification des altérations divergentes entre compartiments pairés

#### Test par bras chromosomique

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

#### Correction multiple

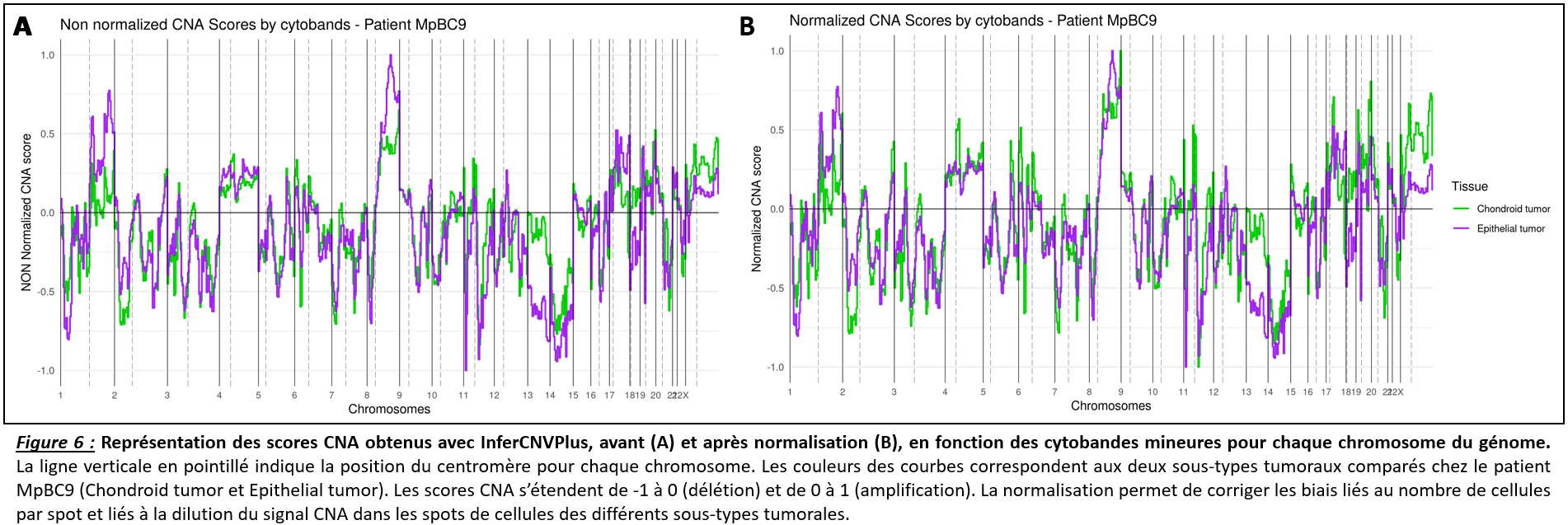
En réalisant les tests statistiques successifs, il est nécessaire de réaliser des corrections multiples pour contrôler le taux de faux positif, dont la probabilité augmente avec le nombre de test réalisé. Dans le cas de notre problématique nous avons appliqué une méthode de correction stricte correspondant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

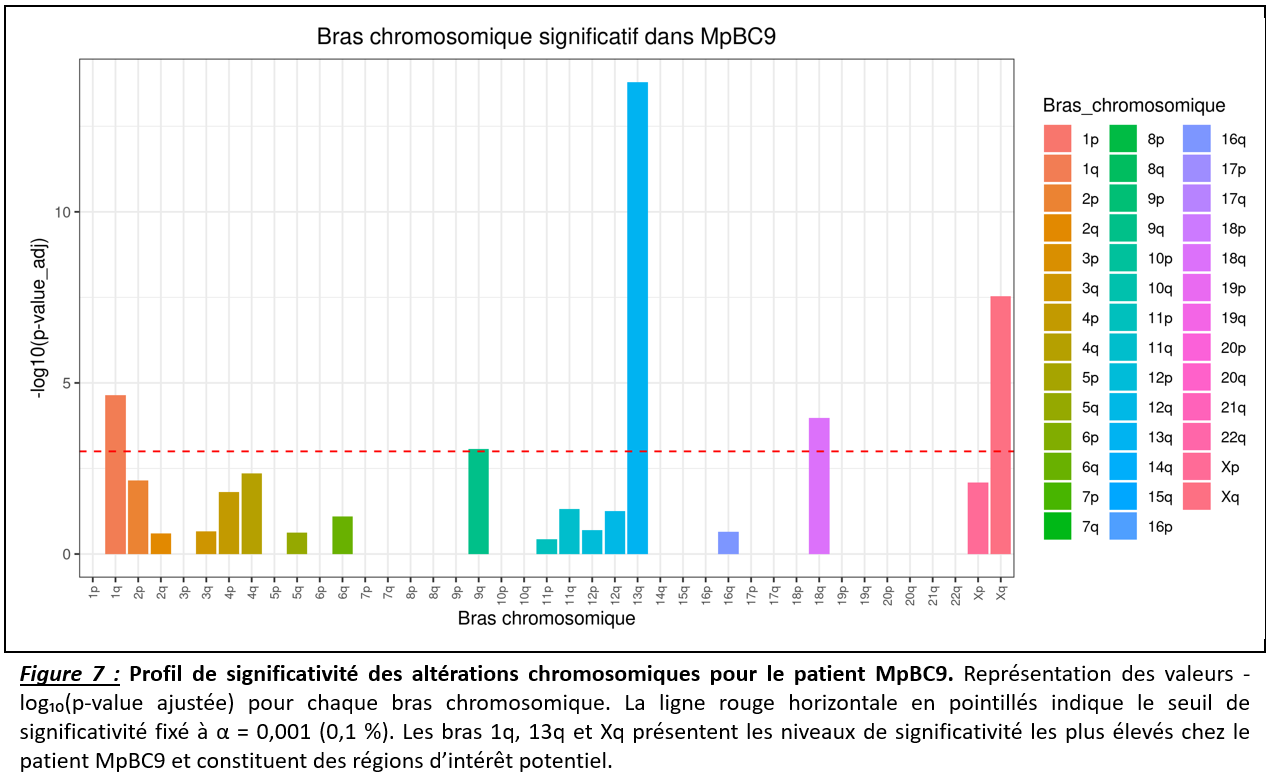
#### Résultats significatifs

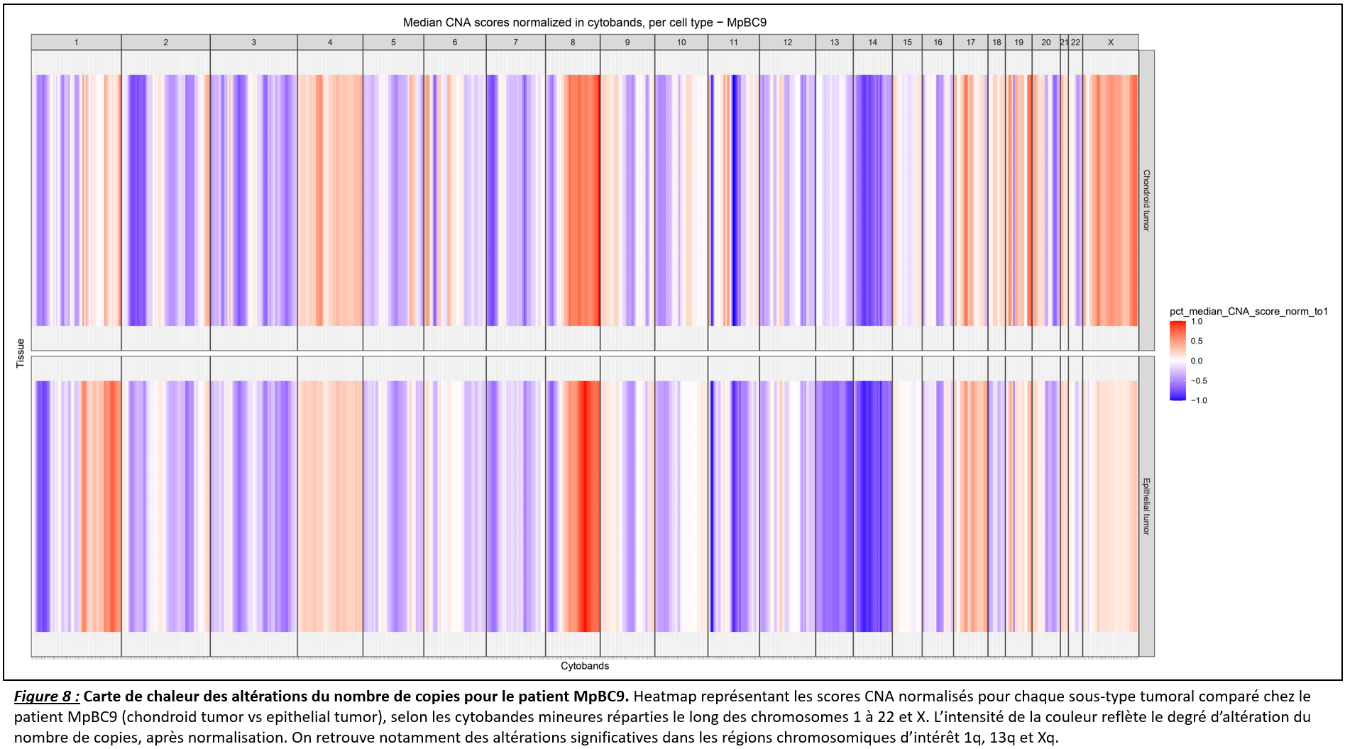
Pour chaque patient, afin de définir si le profil CNA d’un bras chromosomique est significativement différent du profil des autres cytobandes, nous avons utilisé un seuil de 0.1%. En plus de ce seuil alpha, une 2e métrique à été utilisé correspondant à la taille de l’effet. Cette taille de l’effet a été réalisé pour chaque bras chromosique de chaque patient. Elle consiste en la différence des médianes des scores CNA des cytobandes par bras chromosomiques, entre les tissus pairés. Cela permet ainsi de mesurer l’écart des scores médians entre les 2 tissus, pour un bras donné et un patient donné. En effet, en utilisant cette métrique comme filtre, nous pourront éliminer les bras chromosomiques significatifs mais avec une taille d’effet peu importante, et donc moins intéressants.

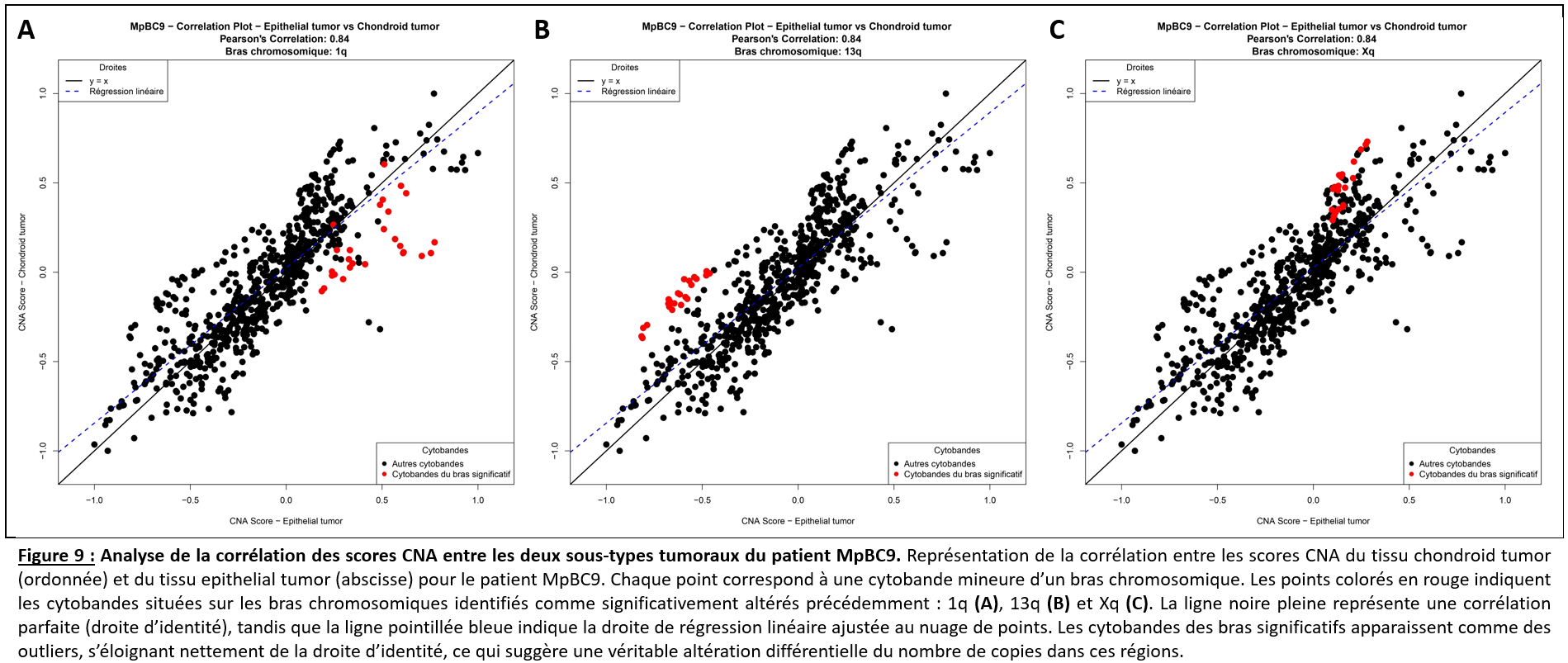
### Visualisation des différences de score CNA entre compartiment

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 CNA 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.

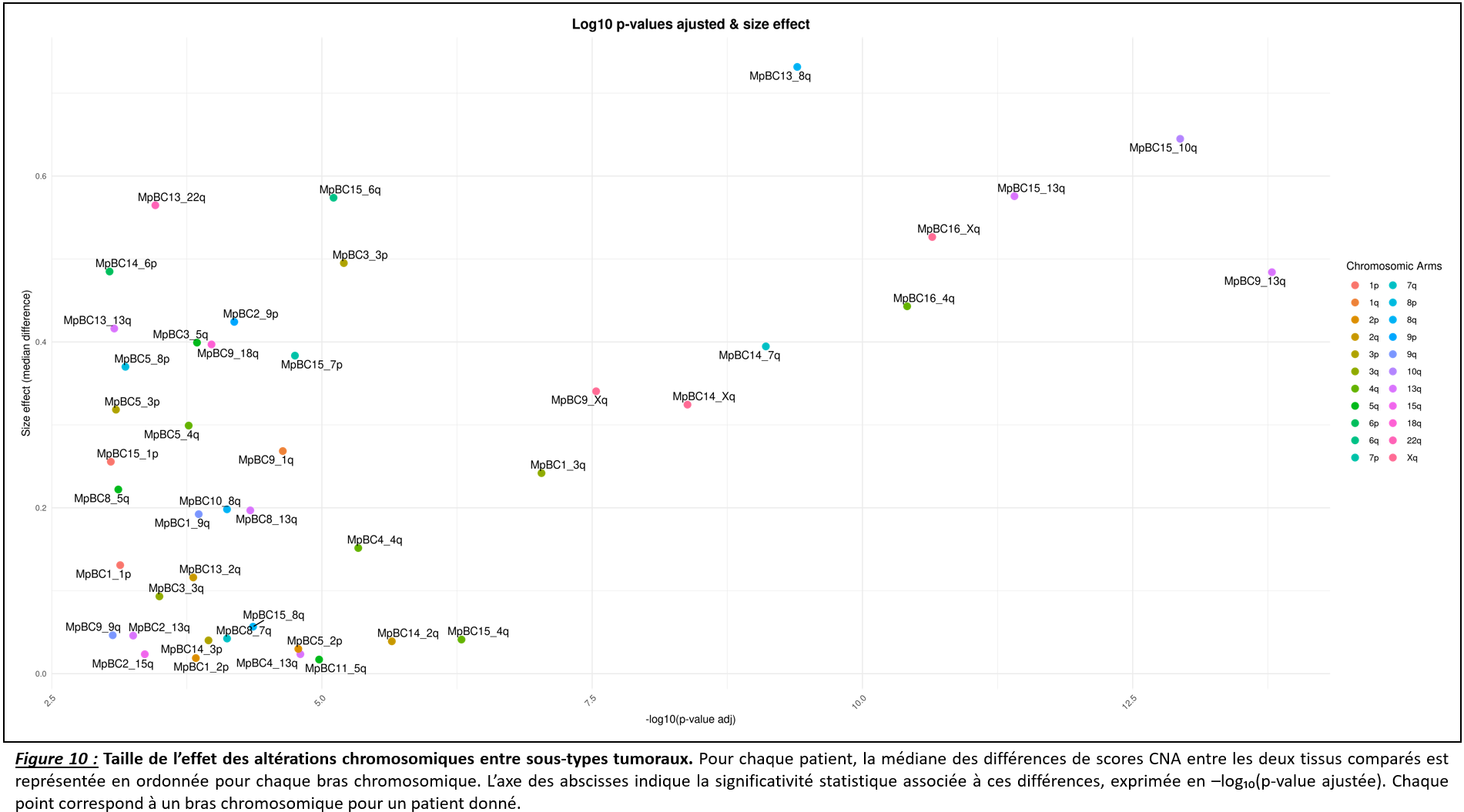


Les tests statistiques réalisé par bras chromosomiques permettent d’identifier les régions chromosomiques significativement altéré comparé aux autres régions du génome chez le patient. On peut ainsi, à partir de la **figure 7**, visualiser les régions génomiques intéressantes et à investiguer. Dans cette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

Une manière 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 : 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 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.



# Discussion

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

### Risque d’hétérogénéité cellulaire intra-spot

#### Séparation imparfaite des clusters (UMAP)

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

La solution permettant de résoudre cette limitation de la technologie Visium et qui permettrait de connaitre la composition cellulaire d’un spot et d’attribuer à chaque sous-type tumoral un profil transcriptomique précis est d’utiliser une technique de déconvolution. Le principe de la déconvolution est de considérer les spots de cellules comme une combinaison linéaire de profils d’expression cellulaire. Ainsi, grâce à cette étape de déconvolution, nous pourrions connaitre les sous-types tumoraux présent dans chaque spot et estimer leurs proportions à partir du profil transcriptomique globale du spot. Des travaux précédents au sein de l’équipe ont déjà montré que des outils, tel que RCTD (Robust Cell Type Decomposition), sont des algorithmes puissants pour la déconvolution de spot Visium et semble plutôt bien adaptés à la biologie des MpBC. Cependant, l’utilisation d’une déconvolution nécessite de comparer le profil transcriptomique mixte de chaque spot aux profils transcriptomiques de référence caractéristique de chaque type cellulaire. Or actuellement, aucuns travaux de recherche n’ont encore construit de tel profil transcriptomique spécifique des différents sous-types tumoraux des MpBC. Je détaillerai plus en détail dans les paragraphes suivants, la manière dont nous construirons ces profils transcriptomiques de référence pour chaque sous-type.

#### Trop faible spécificité des marqueurs phénotypiques

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

Un deuxième axe d’amélioration afin d’identifiés des marqueurs phénotypiques spécifiques des types tumoraux et moins dépendant de l’identité du patient, concerne l’étape de correction du *batch-effect* lors de l’intégration des données transcriptomique de tous les patients. Il est vrai, une correction pas assez importante des comptes d’UMI dans les matrices de comptage de gène des patients peut aboutir à la production de marqueurs plus spécifiques aux patients qu’aux types tumoraux considérés. Le processus d’intégration de multiples de données et de multiples patients est une étape primordiale dans les analyses single-cell, elle n’en demeure toutefois pas moins compliquée. En effet, cette étape délicate nécessite de trouver le bon compromis dans les paramètres entre éliminer les différences propres aux patients sans toutefois sur-corriger les données et éliminer les réelles différences biologiques, au risque de perdre le signal d’intérêt. Dans notre analyse, l’identification de marqueurs spécifiques aux patients (**figure 4B**) suggère que l’intégration des 16 patients et la correction du *batch-effect* avec Harmony n’est pas pleinement satisfaisante et pourrais probablement être amélioré. Ainsi, des ajustements au niveau de la méthodologie ou l’exploration d’outils alternatifs d’intégration pourrait permettre d’affiner encore la robustesse des marqueurs phénotypiques identifiées pour chaque sous-type tumoral.

#### Dilution/Perturbation du signal CNA

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

#### Difficulté pour caractériser des marqueurs génétiques et CNA spécifiques aux sous-types

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

#### Les CNAs divergents identifiés sont-ils passagers ou drivers ?

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

### Consortium MAESTRO : échantillons additionnels

### Analyse de réseaux / Pathways déruglés plutôt que des gènes spécifiques

### Analyse approfondie via snRNA-seq

#### Création d’un atlas MpBC spécifique

### Microdissection des compartiments

#### Déterminants génétiques

##### Mutations drivers divergentes et validation des CNA

##### Epigénétique (méthylome)

### Déterminants non génétiques

#### Déconvolution et composition du microenvironnement

#### Interactions ligand-récepteur tumeur/TME et cœur/bordure

### Validation IHC & Analyses fonctionnelles

# Conclusions

Références bibliographiques

Annexes