Master's degree in Bioinformatics

(Informatique pour la biologie et la santé)

Specialization : Machine learning

2023-2024

Thesis :

Spatial transcriptomics and transdifferentiation in Triple negative breast cancer tumours

Author :

Ms Inès Kardous

Supervised by:

M Pierre Martinez

Internship done at :

Centre de Recherche en Cancérologie de Lyon (CRCL,

UMR INSERM 1052 CNRS 5286, 28 Rue Laennec

69008 Lyon

Université
de Rennes

UFR Sciences de la vie & de l'environnement

ENGAGEMENT DE NON PLAGIAT

Je, soussigné (e) Kardous Inès
Etudiant (e) en Master 1 - BIG

Déclare être pleinement informé (e) que le plagiat de documents ou d'une partie de documents publiés sous toute forme de support (y compris l'internet), constitue une violation des droits d'auteur ainsi qu'une fraude caractérisée.

En conséquence, je m'engage à citer toutes les sources que j'ai utilisées pour la rédaction de ce document.

Signature

INSERM U1242 OSS Equipe
PROSAC

Centre Eugène Marquis Avenue de
la Bataille Flandres Dunkerque
35042 Rennes

Annabelle MONNIER
annabelle.monnier@univ-rennes1.fr

TÉL. 33 (0)2 23 23 61 14

Acknowledgments

I want to express my appreciation to my supervisor Mr Pierre Martinez. During this internship I consolidated and broadened my knowledge thanks to his invaluable insights and expertise. He trusted my work which helped me gain confidence. Pierre also continuously provided encouragement and was always willing and enthusiastic to assist in any way he could throughout the research project. He also contributed in making one of my dreams a reality: presenting my work at a bioinformatic congress.

This academic journey wasn't the easiest and held lots of challenges, but I am very grateful for my family who made this journey less stressful and more enjoyable. I had the best supporters. On top of that I would be remiss if I did not express my gratitude towards my professors who taught science with passion and made me the bioinformatician that I am.

This thesis is for my mom, who showed me unwavering support across my academic years and in my life in general despite her own life challenges. She motivated me to always give my best at work and strive for excellence. She frequently reminded of the importance of acquiring knowledge in all types of discipline. My mom also pushed me to dream big and dare to be ambitious. She is my everything. Words will never be enough for me to express my gratitude towards her.

Last but not least, I want to thank God for putting all these beautiful people in my way. I am who I am and where I am today thanks to all these beautiful souls. Again, I express my sincere thanks to each one of them.

Sincerely,

Inès.

Ps: I didn't expect that I would be this emotional writing those lines.

Summary

Abbreviations

TNBC: Triple Negative Breast Cancer

MpBC: Metaplastic Breast Carcinoma

scRNAseq: Single cell RNA sequencing

SRT: Spatially Resolved Transcriptomics

NGS: Next GEneration Sequencing

Log2FC: Log 2 Fold Change

MAST: Model-based Analysis of Single Cell Transcriptomics

UMI: Unique Molecular Identifier

CAFs: Cancer Associated Fibroblasts

GSEA: Gene Set Enrichment Analysis

NST : Non specific type

Introduction

Breast cancer is the number two killer among cancers in women [1]. During her life, a woman has a 12.9% risk of developing an invasive breast cancer [2]. These past decades, multiple advances in breast cancer treatments efficiency have contributed to enhancing life expectancy but there are still cancers that are difficult to treat[2]. Breast cancers are generally classified according to alteration in three different receptors that can be targeted by specific therapies : oestrogen (ER) and progesterone (PR) receptors, as well as human epidermal growth factor receptor 2 (Her2)[1]. Triple negative breast cancers (TNBC) belong to the subtype characterised by the absence of alterations in all three receptors. It is both a rare and aggressive subtype, representing between 10 and 20% of all breast cancer cases, and is unfortunately still not well characterised yet[3]. TNBC is also characterised by high heterogeneity and plasticity. A recent study also suggests that TNBC tumours have higher rates of proliferation compared to the common subtypes[1].

To date, no biomarkers driving the very specific carcinogenesis mechanism of the TNBC tumours have been identified[1]. As a matter of fact, TNBC is ineligible to targeted therapies and therefore difficult to treat, making it the most lethal breast cancer subtype. Treatments provided are limited to chemotherapy and surgery [1]. Resurgence is also another factor contributing to the high risk of mortality[1]. Triple negative breast cancer patients are more prompt to experience a resurgence of the cancer[1]. Finally, it has been observed that metastasis in TNBC reaches lungs, liver or even brain more often than bone, the typical site for the most common breast cancer subtypes[1]. TNBC is more common in women less than 40 years old and is known for growing and spreading quickly but its symptoms are similar to the common subtypes[4] making it a very challenging cancer to treat.

Cancer cells develop in response to mutations in genes. Only between 1 to 10 percent of breast cancers are inherited, in contrast, the majority is due to acquired mutations [5]. Genes mutations causing cancer are gene alterations that lead to turning on oncogenes or turning off tumour suppressors. Tumour suppressors normally regulate cell division mechanisms like repairing DNA alterations. In the TNBC classification itself, multiple different subtypes have been identified. Subtypes such as metaplastic breast carcinoma (MpBC) are associated with high non-genetic plasticity responsible for high resistance to treatments. MpBC is characterised by an abnormal transdifferentiation of epithelial cells to a different phenotype, whose presence is not expected in the tissue, detected by histopathological review of tumour samples. Previous research has established that metaplastic subtypes “have elevated expression of cell motility pathways, receptor interaction with the extracellular matrix (ECM), and cell differentiation pathways”[1].

Different types of MpBC exist, according to the trans differentiated phenotype observed. These include (but are not limited to) mesenchymal “spindle cell”, squamous, chondroid (cartilage-like), and osteosarcomatoid (bone-like) trans-differentiated cells. The molecular mechanisms behind this plasticity remain however unknown. As different subtypes of MpBC display different survival rates, appropriate and specific molecular markers to complement pathology-based diagnostic options are also still required.

This study is aimed to provide new insights on the transdifferentiation of metaplastic tumours and try to unravel the mechanisms that drive them, using spatial transcriptomics data as well as annotations provided by a pathologist. The primary goal is to analyse the genomic, transcriptomic as well as the microenvironmental differences between the different trans-differentiated compartments, with a specific focus on mesenchymal spindle cell MpBCs. Ultimately, this research seeks to determine the genes and pathways involved in the transdifferentiation mechanism and potential specific biomarkers of these tumours.

Material and Methods

A. Visum 10xGenomics - Spatial gene expression

For the purpose of unravelling the biological process that drives metaplastic transdifferentiation in breast cancer tissues, data collected from 8 MpBC patient samples are studied. To understand both the architecture and the gene expression across our samples, spatially resolved transcriptomics analyses were performed. Spatial transcriptomics approach is performed in order to capture both histological and transcriptional information, as opposed to scRNAseq that doesn't preserve spatial context. The Visium technology provides counts of gene transcripts for each of the 4,992 spots located on a slide. Its use can reveal specific expression features in the mesenchymal tumour cells compared to epithelial tumour cells or normal mesenchymal cells. The use of coverage-dependent methods can furthermore be informative on the genomic differences, in terms of copy number alterations, between these different types of cells in the same patient.

The tissue slides used in this study were formalin-fixed paraffin-embedded (FFPE). FFPE technique preserves tissues at the expense of RNA, however 10xGenomics provides a method using FFPE that ensures a good quality gene expression readout[6]. Each Visium FFPE capture area contains 4992 barcoded spots corresponding to an average resolution of 1 to 10 cells, and can profile the expression of 18,085 genes.

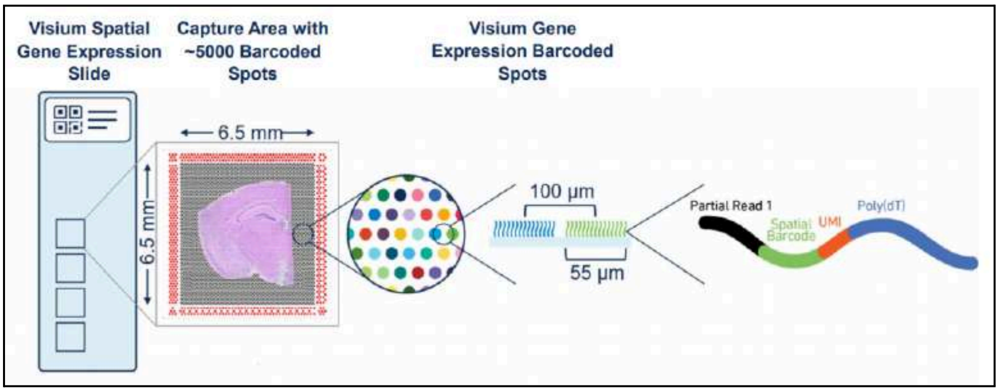

Visium Spatial Gene Expression Slide
Capture Area with ~5000 Barcoded Spots
Visium Gene Expression Barcoded Spots
6.5 mm
6.5 mm
100 µm
55 µm
Partial Read 1
Spatial Barcode
UMI
Poly(dT)

Figure 1 : 10xgenomics Visium technology
Reference : https://ngisweden.scilifelab.se/methods/10x-visium/

The Visium method consists of in situ capturing of RNA and ex situ sequencing enabling a trade off in RNA sequencing resolution coverage and depth and spatial resolution. SRT enables a detailed characterization of coordinated multi cellular processes and the possible influence of immune composition. The biological sample is placed on a capture area with barcoded spots (Fig.1). RNA molecules are then released and captured. Since our samples are FFPE, polyadenylated mRNA are captured. The complements of the barcodes are then incorporated and the cDNA libraries are generated in order to sequence using Next generation method (NGS) and get the gene expression profiles.

B. Data preprocessing - SpaceRanger and LoupeBrowser

The current study uses data that was upstream produced using Visium 10xGenomics technology and preprocessed with the SpaceRanger 2.0.0 software provided by 10xGenomics. SpaceRanger yields quality control (QC metrics), a barcode-feature (or gene-spot) matrix, and a .cloupe file for interactive use with the LoupeBrowser visualisation and analysis software developed by the manufacturer. Reads were demultiplexed and mapped to the GRCH38-2020-A human reference genome. A gene-spot matrix was produced by computing the transcripts counts in each spot using the barcodes [7]. Prior to this internship, combinations of 1 to 10 clusters were determined using K-means clustering from the LoupeBrowser software. Final clusters were defined, reviewed and annotated manually for each sample, with the help of a trained pathologist. Following the results of subsequent analysis (biomarker identification in particular), annotations for patient MpBC5 data were revised and all analyses were updated.

C. Cell-type deconvolution

Visium data is not as precise as single-cell resolution technologies, and multiple cells (~1-10) are generally present in the space covered by a single spot. In other words, in each spot several different cell types can be found. That's why cell-type proportion is an important way to describe the genuine content of a spot. In order to leverage the relative proportion of each cell-type in the counts computed, deconvolution has to be performed.

During this internship, several tools were tested. Some failed to be installed such as Stereoscope due to several versioning errors. Deconvolution of the Visium spatial transcriptomics data was first performed using SpaDecon v1.1.2[8]. SpaDecon is a semi-supervised tool that aims to overcome the insufficient resolution in our spatial transcriptomic data, not only using gene expression data but also spatial and histological information, in order to evaluate the cell type proportions in each spot[9]. Annotated scRNA-seq gene expression data from the same type of tissue as the SRT data are required for deconvolution. We used the same training set for SpaDecon, as the original authors: the reference Breast Cancer Cell Atlas, containing 100,064 cells and 9 cell types, from 30 breast cancer samples[10].

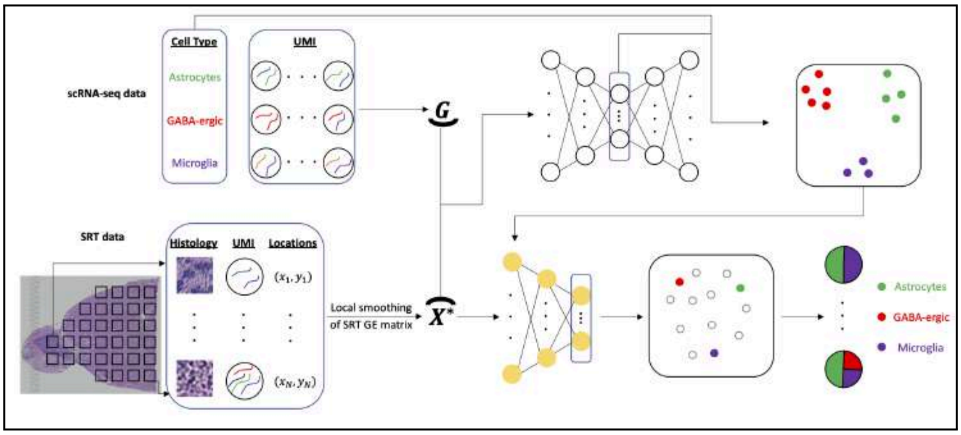


Figure 2: SpaDecon deconvolution process

Reference: https://github.com/kpcoleman/SpaDecon

A conda environment was created in order to install the python 3.7 version required to run SpaDecon. Prior to the install of SpaDecon some python packages were required : keras 2.2.4, pandas 1.2.4, numpy 1.20.1, scipy 1.6.2, scanpy 1.7.0, anndata 0.7.6, sklearn, tensorflow 1.14.0. SpaDecon is performed on the SRT (includes spatial and histological information) and scRNAseq data (Fig.2). SpaDecon analysis could be performed using both the spatial and histological information only on two patients (MpBC6 and MpBC7) but failed on the others. SpaDecon was not optimised with the more recent FFPE protocol for Visium data, and we thus decided to only rely on spatial and transcriptomics data for SpaDecon-based deconvolution. Outputs were visualised on R using Seurat library[11]. Finally, the RCTD tool was also used, from the R package spacexr[12]. RCTD, as opposed to SpaDecon, relies on only part of the reference file in order for it to run efficiently (efficient in time and memory). The normalize_weights() function normalises the weights obtained so that in each spot the sum of the cell type proportions is equal to 1 for easier comparisons and so interpretations. After reviewing the respective results of each method, we decided to pursue analyses with RCTD.

D. Analysis of immune infiltration of the spots

In this step, we looked at the histological annotations provided by the pathologist and compared it with the deconvolution cell type profiles obtained for each spot. In multicellular processes, it is interesting to look at the possible influence of immune composition. Results were plotted as bar plots using the ggplot R package on all patients' data pooled together (per patient barplots were also done, but because of a lack of space we judged better to present the general trend). The relative cell types proportions were then compared using anova in order to determine if the differences in cell infiltration observed between the spots annotated by the pathologist as mesenchymal tumour spots, normal mesenchymal spots, squamous and epithelial tumour spots were significant. A post-hoc Tukey HSD test was then conducted in order to detect which comparisons are insignificant with p-values < 0.001. In the barplots, we grouped the pathologist annotations the following way :

- Epithelial tumour cell annotations: Epithelial tumour cells, NST (Non specific type) cells, NST surrounded by spindle, Epithelial tumour;
- Mesenchymal tumour cell annotations: Mésenchymal tumour cells, Spindle surrounded by NST, Spindle cell tumour, Spindle cell tumour, Spindle cell tumour, Spindle++ spindle, Spindle- tumour, Mixoid matrix-enriched spindle+ spindle, Classic spindle tumour;
- Normal Mesenchymal cell annotations: Non tumoural fibrous tissue, Normal fibrous tissue;
- Squamous tumour cell annotations: Squamous- tumour, Squamous cell tumour, Squamous++ tumour.

This is one of the analyses that had to be repeated on the updated data in order to determine the possible consequences of the MpBC5 annotation error.

E. InfercnvPlus - Copy number alterations

Relative copy number alterations (CNA) scores were computed using the infercnvPlus R package[13]. InfercnvPlus is an enhanced version of the package Infercnv that is suited for spatial transcriptomics data. For each patient the CNA scores were computed using as a reference the cells annotated as : blood, non tumoural fibrous tissue, immune cells, normal fibrous tissue, normal fibrous stroma. In our dataset, some patients do not have normal cells. Hence, we included in each calculation the normal cells of all patients. That way, each patient (even those that have normal cells in their data) have the same reference cells and results can be compared. Additionally, we removed the data of the cells we did not study, such as necrosis or adipose cells. Individual CNA profiles of genes were averaged on a per (minor) cytoband basis, according to the UCSC genome browser Hg38 cytoband reference file[14]. For each spot, a positive CNA score suggests a gain in genomic material compared to our reference cells, when a negative value means a loss in genomic material.

We seeked to detect genuine CNA between the different compartments of each tumour. Data were normalised by computing z-scores for the difference of the CNA scores between different compartments of each individual tumour. We then applied Bonferroni multiple-testing correction to detect the significant outliers. We investigated if these outlier alterations (most likely to be genuine gain or losses of genomic loci) were contiguous, using a heatmap representation. This analysis was redone with the updated MpBC5 annotations.

F. Differential gene expression analysis - Finding biomarkers specific to the mesenchymal tumour compartments

We investigated the specificities of gene expression between the different tumoural and stromal compartments using the Findmarkers[15] function from the Seurat R library**.** The function is built to determine gene differential expression by comparing two sets of data and giving as an output an adjusted p value as well as the Log 2 Fold Change (Log2FC) metric. In order to visualise the results of our analysis we plotted the outputs on Volcano plots by using EnhancedVolcano R library [16]. Significantly differentially expressed genes were defined using p value < 0.05 and an absolute log2 fold change > 1. The figures being too difficult to read, we then tried to apply more stringent parameters in order to define premium differentially expressed genes that could be used as molecular markers to identify mesenchymal tumour cells and better categorise TNBC. We first pooled all the patients data and performed a global comparison between (1) the mesenchymal tumour spots and (2) the union of epithelial or squamous tumour spots and normal mesenchymal ones. We filtered and kept only the genes with p-values < 0.001 and Log2FC >= 1. This was designed to identify markers extremely specific to the mesenchymal trans differentiated tumour cells. Findmarkers found more than 1,300 potential biomarkers, which is too many genes to be considered as biomarkers and made us question the performance of this tool…

The specificities of gene expression between the different tumoural and stromal compartments was then investigated using the Model-based Analysis of Single Cell Transcriptomics (MAST) R package[17]**.** In this study, we were interested in finding molecular markers to identify mesenchymal tumour cells and better categorise MpBCs. For biomarkers to be relevant, they have to be common to the majority of patients. As a matter of fact, all patients' data was pooled together (Some patients have no mesenchymal tumoral cells such as patient MpBC8 but do have epithelial tumoral cells). A global comparison was performed between (1) the mesenchymal tumour spots and (2) the union of epithelial or squamous tumour spots and normal mesenchymal ones.

The specificity of MAST relies on the broad possible options when performing the analysis. In the zero-inflated regression model (zlm), it is possible to precise both fixed and random effects. In order to get rid of the possible errors induced by the use of all patients' data pooled together, a random effects variable on patients' identifiers was added. As for the fixed effects, cell annotations and total Unique Molecular Identifiers (UMI) counts were used. Outputs were filtered based on Bonferroni-adjusted p-value < 0.001 and Log2FC >= 1. MAST computes log2FC assuming that the data has been priorly log transformed. A focus was made on the genes that were upregulated in the mesenchymal tumour compartment since a biomarker gene is more easily detected when upregulated than when silenced, especially with cross-technological validations such as histology. This was designed to identify markers extremely specific to the mesenchymal trans-differentiated tumour cells.

The second step consisted in the selection of a few high-quality biomarkers. We decided to keep the genes that are significant when we complete all pairwise comparisons between (1) the mesenchymal tumour and epithelial tumour and (2) between the mesenchymal tumour and the normal mesenchymal compartments of patients MpBC1, MpBC2, MpBC3, MpBC5 and MpBC6 (only patients having both mesenchymal tumour compartments and epithelial tumour cells, squamous tumour cells or normal mesenchymal ones).

We proceeded the following way. We compared (1) the mesenchymal tumour compartments with the epithelial tumoural ones of patients MpBC1, MpBC2, MpBC5, (2) the mesenchymal tumour compartments with the normal mesenchymal one for patients MpBC1, MpBC2, MpBC5 and MpBC6 (3) the mesenchymal tumour compartments with the squamous one of patient MpBC3. First we removed the genes that have a log2FC < 0 with a significant p-value (<0.001) in at least one individual comparison, to ensure a global overexpression trend for each biomarker. The next step consisted in filtering in order to keep top biomarkers of the mesenchymal tumour compartments. We kept the genes that have log2FC >= 1 and p-value < 0.001 in at least 5 individual comparisons. Log2FC was computed by computing the difference between compartments of the average logged UMI counts. The p-values were retrieved from the Bonferroni adjusted p-value results of t-tests. The Vlnplot function of the Seurat package was used to plot the results of all comparisons for all genes of interest.

The initial results indicated that most high-quality biomarkers were not significantly differently expressed in sample MpBC5. This prompted the re-examination and subsequent redefinitition of this patient’s annotated clusters. This biomarker analysis was then performed de novo with the updated annotated data.

G. Gene set enrichment analysis (GSEA)

This step consists in identifying sets of genes that are enriched in particular differential expression analysis output, to identify changes across pathways. It determines if gene sets show significant or concordant changes. A GSEA tutorial was followed [18] and fgsea R package was used [19]. Gene ontology and reactome gene sets were used as references. In this step, as opposed to the search for biomarkers, less stringent parameters were applied and both up- and down- regulated genes were kept. For this step we applied thresholds of pval.adj <= 0.001 & abs(log2FC)>= 0.5 on the MAST model that compared, across all patients, the mesenchymal tumour compartment with the reference spots (reference includes : epithelial tumour cells, squamous tumour cells and normal mesenchymal cells). We decided to keep the pathways with a p-value < 0.01.

Results

A. Deconvolution

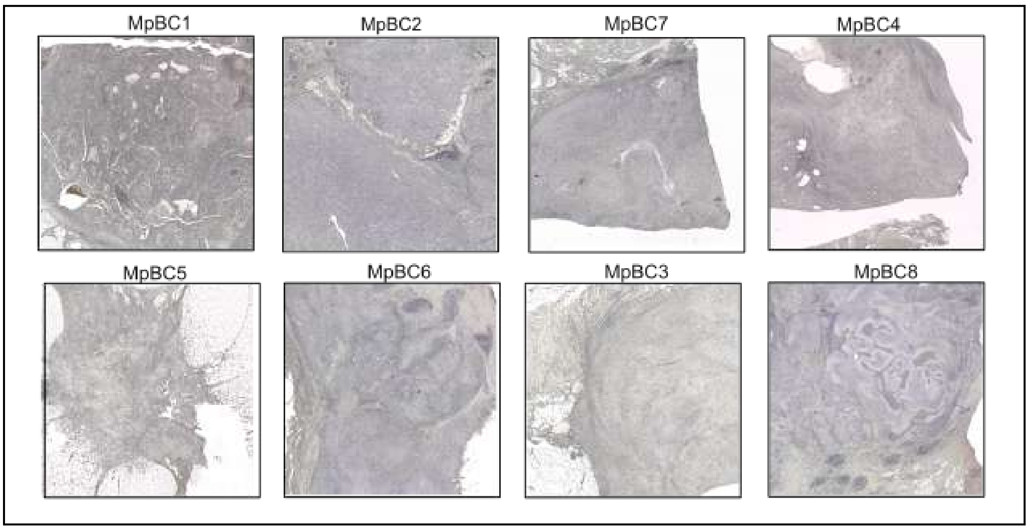


Figure : Raw patients slides
These are the areas of each tumour that were selected to be studied with the SRT.

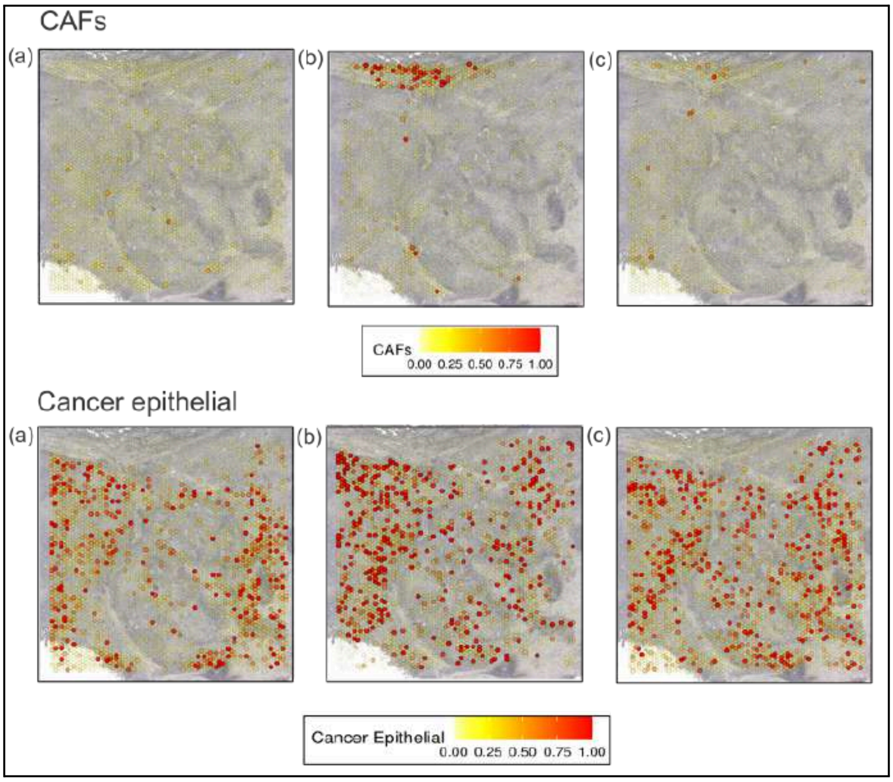


Figure 4: Example of MpBC6 patient data deconvolution using SpaDecon

(a) Results obtained when using only the spatial data along with the scRNAseq
(b) Results of the deconvolution done using only the histological data along with the scRNAseq
(c) Results of the deconvolution done using both the spatial and histological data along with the scRNAseq

We performed deconvolution in order to overlay tissue structure (Fig.3) with cellular composition, and thus better understand the microenvironment organisation of mesenchymal breast cancer tumours. Thanks to the deconvolution, we can visualise and determine cell-types present in the slides and consequently analyse the microenvironment.

Deconvolution results obtained with SpaDecon using only the spatial data are for the CAFs and epithelial cancer cells closest to the results obtained when using both spatial and histological information (Fig.4). We can notice that there are very few to almost no CAFs (cancer associated fibroblast) found with SpaDecon. This is surprising since patient MpBC6 presents mesenchymal features both in their morphology and transcriptomic profile, with CAFs being the only mesenchymal cell type in our training data set (breast cancer single cell atlas). We also observed numerous spots with high scores for the cancer epithelial cell type, which is not supposed to be present in this pure spindle cell MpBC sample. .

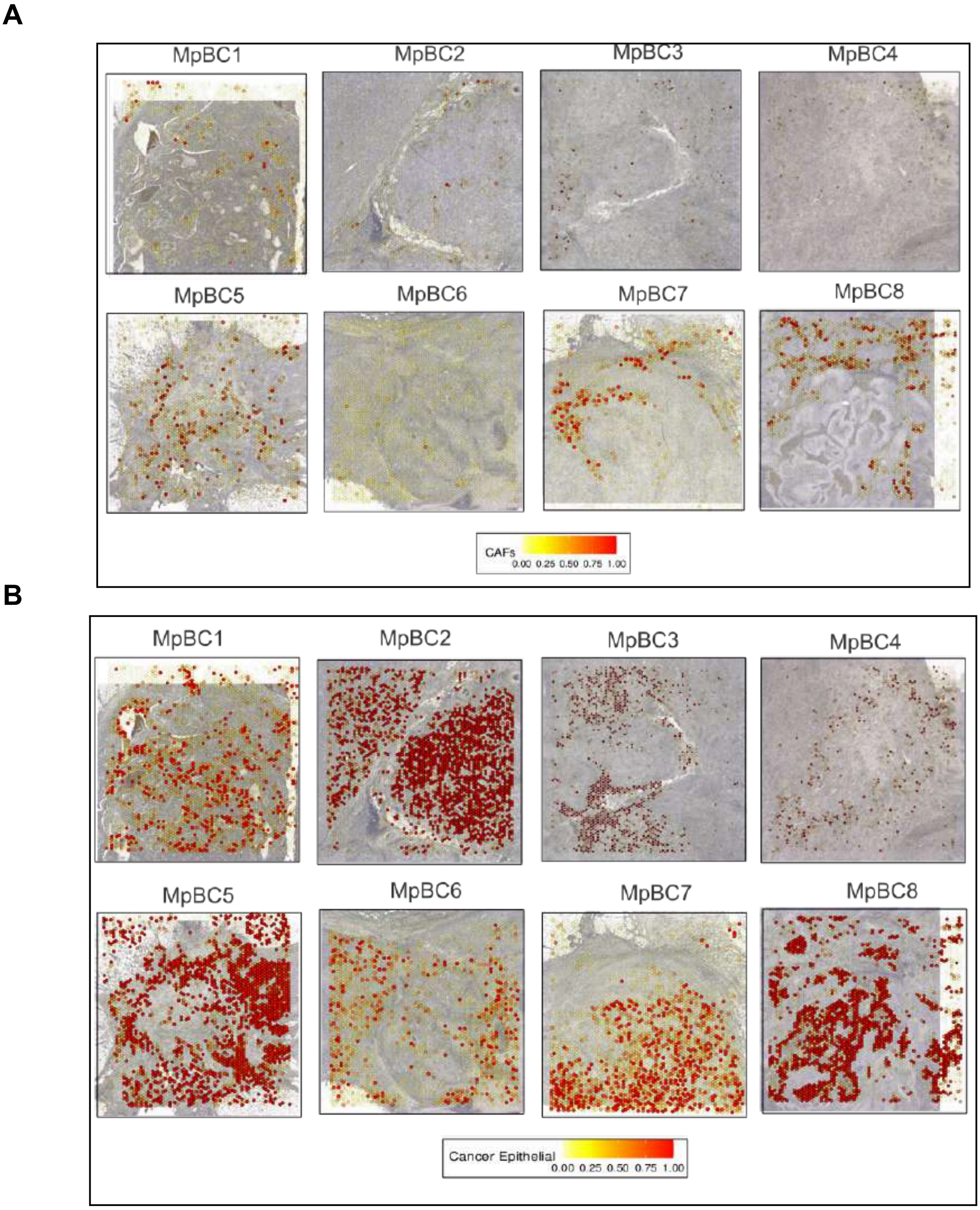
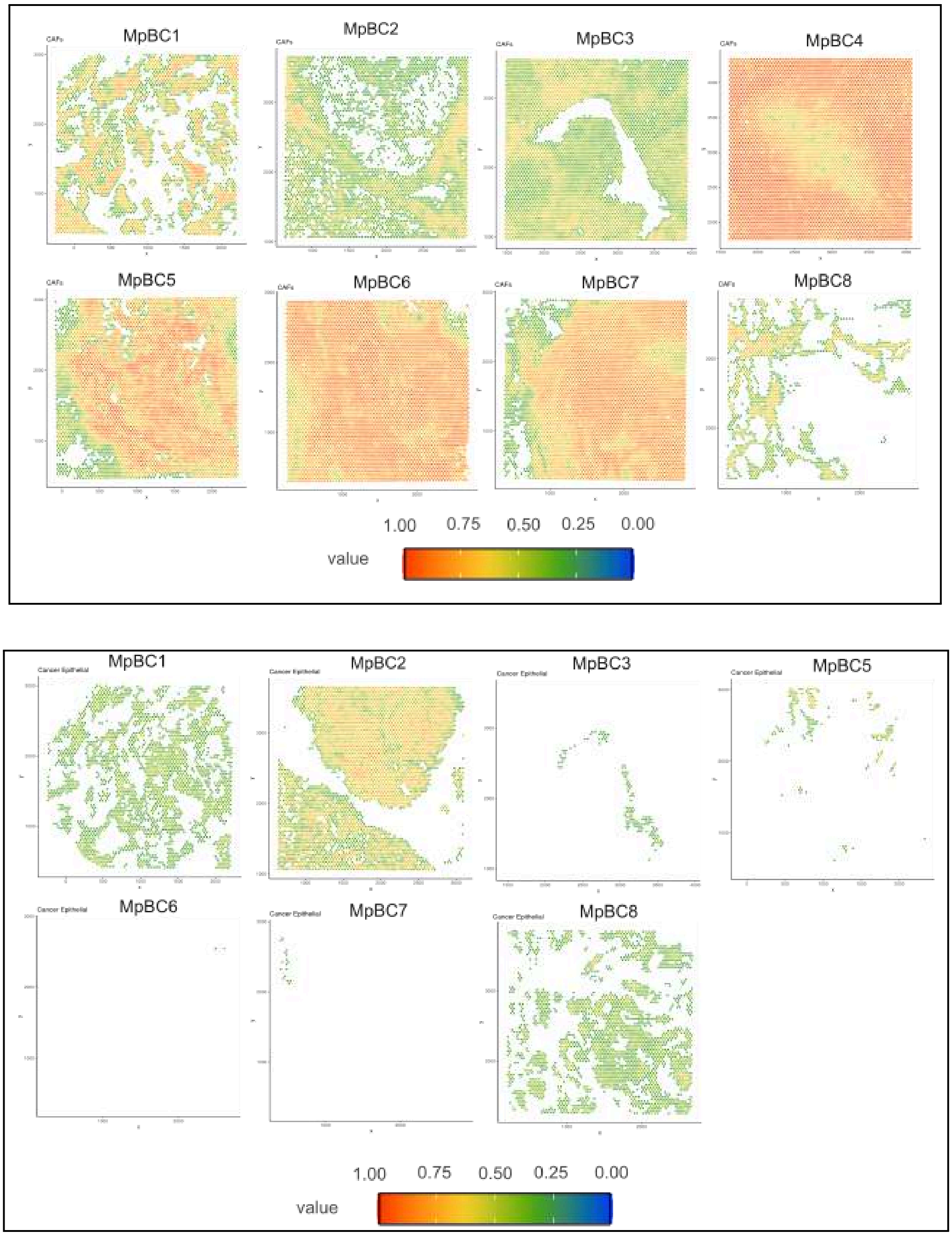


Figure 5: SpaDecon deconvolution outputs - Overlaying deconvolution in spatial transcriptomics data
(A) CAFs Results obtained for the deconvolution done using the spatial data of the patients along with their scRNAseq data. Since the atlas used as a reference for the deconvolution doesn't include mesenchymal tumoral

cells, Mesenchymal tumour cells are histologically and transcriptionally closest to Cancer associated fibroblasts (CAFs). However there are no normal mesenchymal tumour cells in the atlas either, which means CAFs can also correspond to normal mesenchymal tumour cells. **(B) Cancer epithelial** Results obtained for the deconvolution done using the spatial data of the patients along with their scRNAseq data.

Following the results obtained for patients MpBC6 (Fig. 4), the deconvolution was performed using spatial information on the remaining patients (Fig.5). SpaDecon rarely found any CAFs, in most of the patients. We can notice on patients MpBC5, MpBC7 and MpBC8 that the CAFs detected are spread across the slides without any obvious spatial pattern. In addition, patient MpBC7 seems to present cancer epithelial cells when according to the pathologist there are no epithelial cancer cells, but only mesenchymal spindle cell tumour cells. Similarly, patient MpBC8 doesn't present cells that are histologically similar to mesenchymal normal or tumoral cells but SpaDecon detected CAFs during the deconvolution. Altogether, these results suggest a poor performance of the SpaDecon tool in our data.



A

B

Figure 6: RCTD deconvolution outputs - Overlaying deconvolution in spatial transcriptomics data
(A) Cancer associated fibroblasts (CAFs) Weight values are represented in gradients of colours, blue for 0 and red for 1.The sum of all proportions is equal to 1. **(B) Epithelial tumour cells** Weight values are represented in gradients of colours, blue for 0 and red for 1.The sum of all proportions is equal to 1. MpBC4 does not present any spot categorised as epithelial tumour cells when looking at the confidence thresholded results.
The white regions correspond to regions that were removed by RCTD because they do not respect the confidence threshold determined by the authors of the tool. The confidence threshold (by default 5) corresponds to "the minimum change in likelihood (compared to other cell types) necessary to determine a cell type identity with confidence"[20]

As opposed to SpaDecon, RCTD appeared to be much more performant. RCTD better detects CAFs (Fig.6A). Additionally, as opposed to SpaDecon there seem to be barely any epithelial tumour cells in patient MpBC6 or MpBC4 (Fig.6B) which is as expected.

The spots displayed in Figure 6 correspond to the cell type identity determined with strong confidence with RCTD. All slides display spots with at least between 25-50% CAFs (Cancer associated fibroblasts) (Fig.6A). Importantly, this is expected since the RCTD training set did not contain trans-differentiated MpBC cells. For this reason, in our data CAFs are often overlapping with mesenchymal normal or tumoural spots, which are the closest match in terms of RNA expression profile. Pathologist annotations help us differentiate between normal mesenchymal cells and spindle cell tumour cells, which are both scoring high for CAF proportion. Epithelial tumour cells and CAF display spatially exclusive expression in most samples. A very intermingled region of spindle cell and epithelial tumour cells is discernible in sample MpBC2 (bottom left quarter), where Visium spot resolution is not precise enough to precisely delineate the two contingents. Additionally, patient MpBC4 presents absolutely no epithelial cancer cells and Slides MpBC3, MpBC5, MpBC6, MpBC7 present barely any epithelial tumour cell (Fig.6B). We thus continued our analyses of the cellular composition of MpBC compartments using the output of RCTD, and discarded the results from SpaDecon.

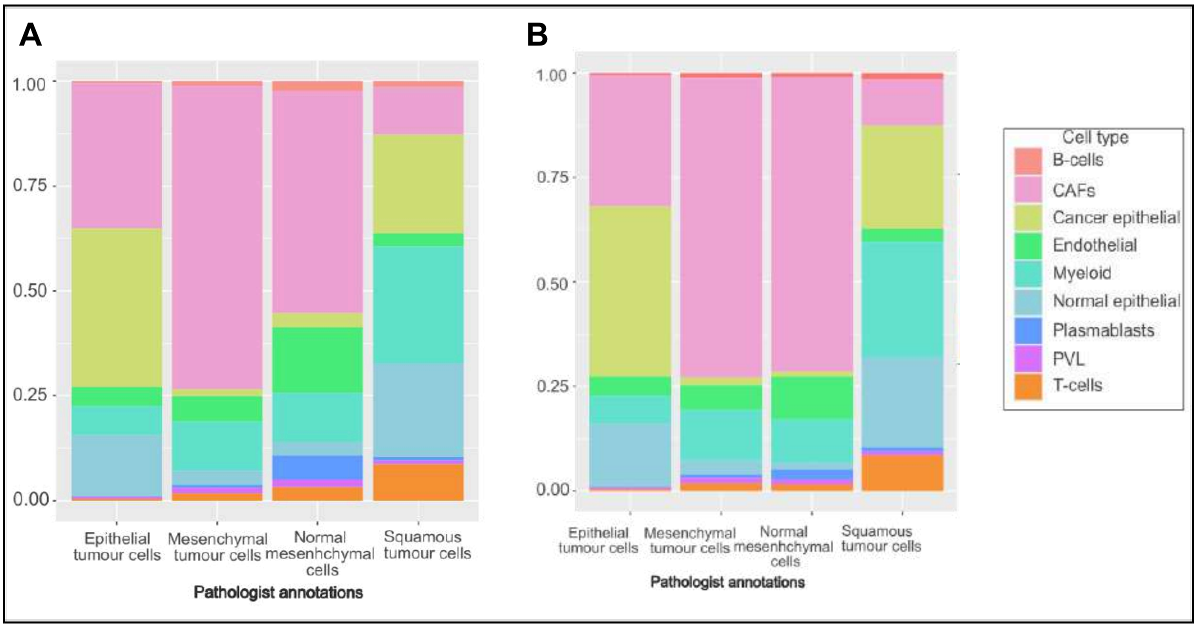


Figure 7: Spots infiltration using RCTD deconvolution outputs
(A) Spots infiltration with the incorrect annotations of patient MpBC5 data (B) Spots infiltration with the corrected annotations of patient MpBC5 data. The spots infiltration corresponds to the results obtained when all patients data were pooled.
Majority of the comparisons are statistically significant in the barplot (A). In the barplot (A), the only comparisons that are not statistically significant are between the normal and tumoural mesenchymal cells in terms of 1/ cancer

epithelial cells proportions (p-value = 0.013 > 0.001), 2/ myeloid proportions (p-value = 0.999 > 0.001), 3/ and lastly in terms of normal epithelial proportions (p-value = 0.997 > 0.001.
Majority of the comparisons are statistically significant in the barplot (B). In the barplot (B), the only comparisons that are not statistically significant are between the normal and tumoural mesenchymal cells in terms of 1/ CAFs proportions (p-value = 0.009 > 0.001), /2 cancer epithelial cells proportions (p-value = 0.062 > 0.001) 3/ and lastly in terms of PVL proportions(p-value = 0.079 > 0.001).

We further explored the spots' cell compositions (Fig.7). Most of the comparisons between the compartments are statistically significant whether we look at the results obtained with the corrected or not MpBC5 annotations. Only some of the comparisons between normal and tumoural mesenchymal compartments differ. When including the corrected data of MpBC5, the proportions of myeloid, normal epithelial become significantly different between the mesenchymal tumour compartments and the normal mesenchymal ones, when PVL and CAFs proportions differences become not statistically different. It appears that squamous tumour cells are the most infiltrated by immune cells such as B and T lymphocytes or myeloid cells (Fig.7A). Squamous tumour cells are epithelial tumour cells that are extremely differentiated. The Mesenchymal normal and tumour compartments are mainly made of CAFs as expected. Additionally, the mesenchymal tumour compartments present barely any immune cells compared to the normal mesenchymal compartment (Figure.7B). Lastly, epithelial tumour spots are significantly more infiltrated by CAFs than the squamous tumour spots.

B. Copy number alterations (CNA)

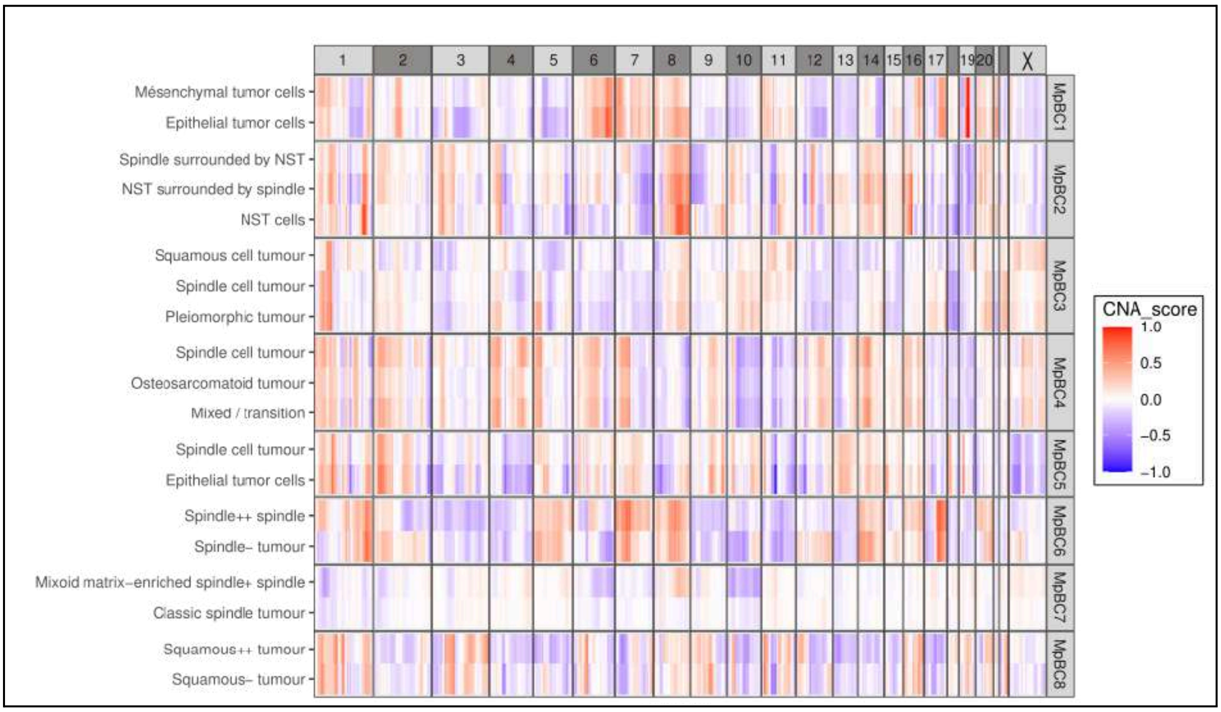


Figure 8: CNA scores in cytobands per cell types
In each cytoband the median CNA score is represented.

There are copy number alterations in almost all compartments compared to the reference cells (blood, non tumoural fibrous tissue, immune cells, normal fibrous tissue, normal fibrous stroma) (Fig.8). Differences between patients can be identified through visual inspection, but are less clear within patients. Patient 2 for instance seems to present differences in DNA copy alterations in its chromosome 9 between the spindle surrounded by NST & NST surrounded by spindle and the NST cells. Patient 3, has slight differences between the spindle & pleiomorphic tumour cells in CNA scores with the squamous tumour ones such as in the chromosomes 1, 5 and 14. Lastly, patient 6 presents differences between its two contingents in its chromosomes 1, 2, and 12 when patient MpBC7 not only does not present differences in CNA between its cell compartments but even with the reference cells which is quite surprising. All in all, this heatmap doesn't present striking genomic differences in terms of copy number alterations.

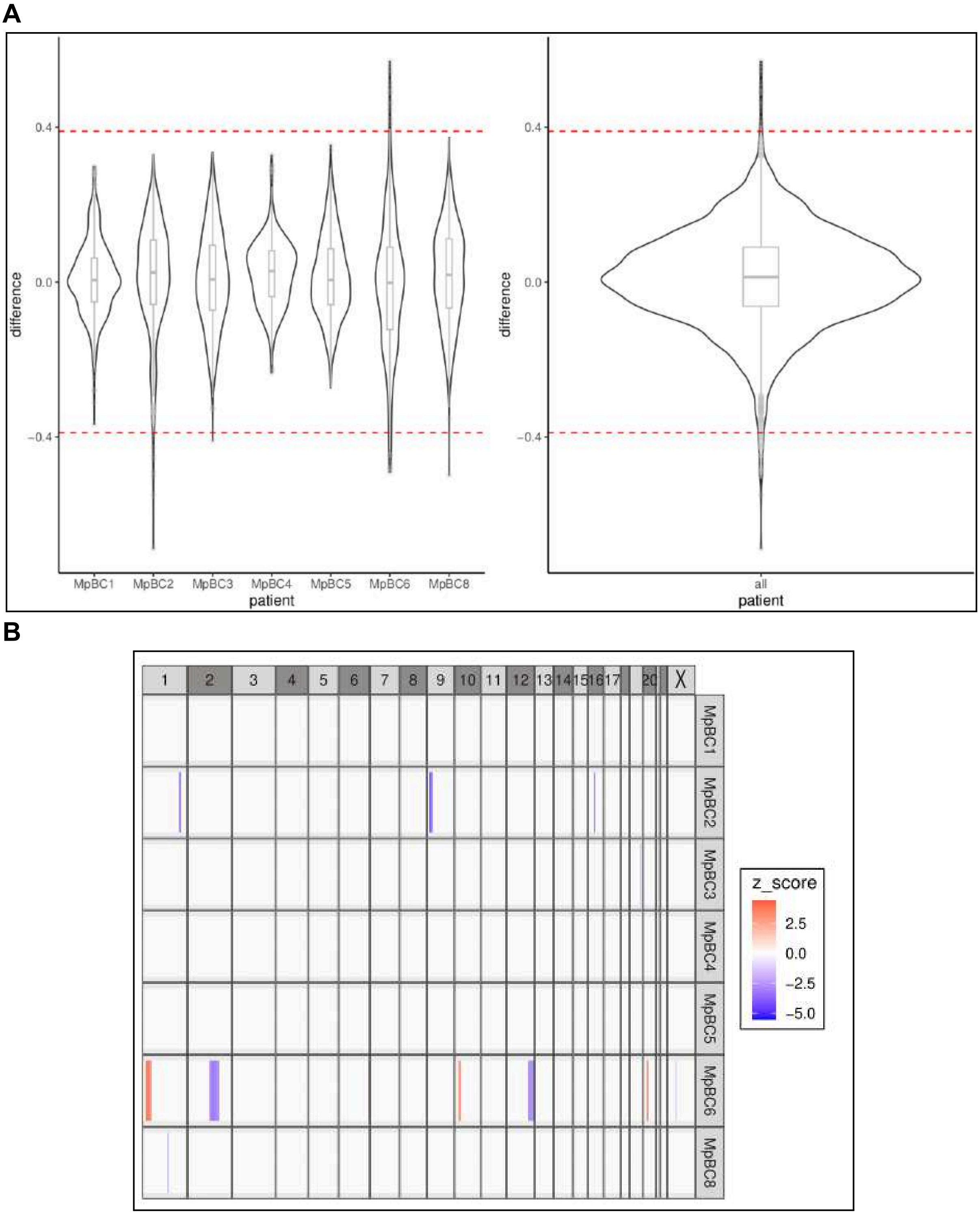


Figure 9: Copy number alterations between compartments
(A) Differences in CNA scores per cytobands computed between two compartments, (B) Identification of significant genomic copy alteration differences(z-scores)
The red line corresponds to the limit CNA score corresponding to a z-score of 3 or -3. The threshold is set at 0.39.

The following differences between compartments were studied: MpBC1 : Mesenchymal tumour cells and Epithelial tumour cells; MpBC2 : Spindle surrounded by NST and NST cells; MpBC3 : Squamous cell tumour and Spindle cell tumour; MpBC4 : Spindle cell tumour and Mixed/Transition; MpBC5 : Spindle cell tumour and Epithelial tumour cells; MpBC6 : Spindle++ spindle and Spindle- tumour; MpBC8 : Squamous++ tumour and Squamous- tumour. These violinplots were done using the corrected MpBC5 data.

In order to see if there were significant outliers in terms of differences between compartments of the same tumour, we computed z-scores on the pairwise differences in CNA scores. It appeared that CNA score difference outliers were detected in patients 2, 6 and 8(Fig.9A). . An abs(z-score) of 3 means that the values of the differences observed between two compartments are at least three standard deviations greater than the average. Interestingly, when plotting on the heatmap the cytobands with z-scores > 3 or z-scores < -3, we can observe that outliers from patient 6 formed contiguous segments (Fig.9B). This strongly indicates outliers. This could indicate the existence of *bona fide* CNA differences between the different compartments in this MpBC patient. The other alterations were deemed too small to be reliable.

C. Gene set enrichment analysis

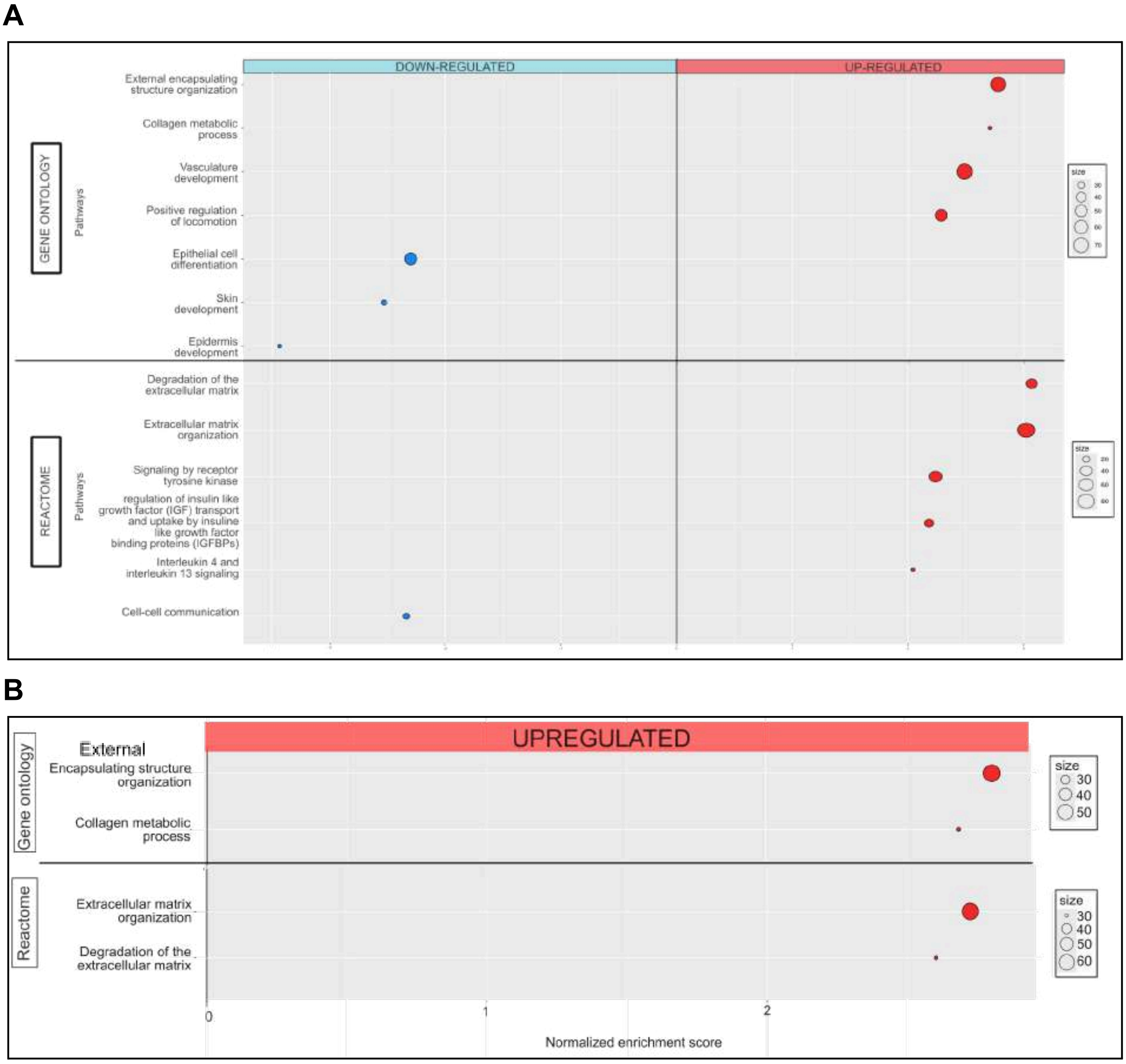


Figure 9: GSEA results

(A) GSEA results obtained using differentially expressed genes obtained with the MAST model run on the pooled data that includes the incorrectly annotated MpBC5 data, (B) GSEA results obtained using differentially expressed genes obtained with the MAST model run on the pooled data that includes the correct annotated MpBC5 data.

GSEA analysis was performed in order to understand which biological pathways are significantly different between mesenchymal tumour cells compartments and the other ones (including only epithelial tumour cells, squamous tumour cells and normal mesenchymal cells). Both the output obtained with the GO (Gene ontology) Reactome references present a majority of up-regulated pathways in the mesenchymal tumour compartment compared with the reference (Fig.9A). The up-regulated pathways are mostly related to extracellular matrix remodelling. On the other hand the processes that are down-regulated are linked to the development of epidermis or cell-cell communication (Fig.9A). However, the results of Figure 9A were obtained using the incorrect MpBC5 annotations, and were thus updated. In fact, when looking at the GSEA output that used the correctly annotated MpBC5 data (Fig.9B), we can notice that both GO and Reactome detect only differentially upregulated pathways. The pathways found are related to external matrix organisation, external structure organisation and collagen metabolism. This shows us how annotation errors can impact the results by including several potential false positives, which can be misleading when interpreting the biological mechanisms driving MpBC transdifferentiation.

D. Biomarkers

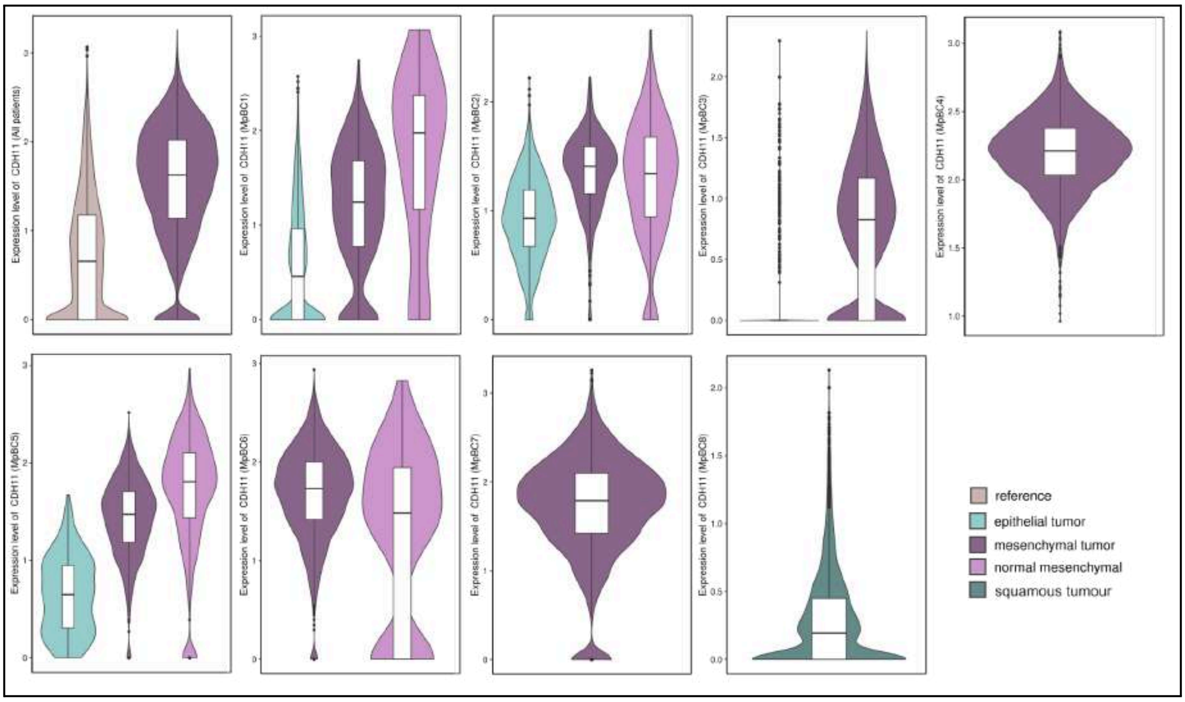


Figure 10: Example of the CDH11 gene's expression levels per patients (and per compartments)

At first we kept all the genes that have an expression level that is globally superior in the mesenchymal tumoural spots compared to the others compartments. However, we implemented further control to identify high-quality genes that are recurrently overexpressed in individual mesenchymal tumour compartments (see Methods). For instance, despite its significant p-value in the overall comparison, the "CTH11" gene was not deemed satisfactory (Fig 10). In fact CTH11 doesn't seem to be a good mesenchymal tumour biomarker since in patient MpBC1 or MpBC5 the level of expression is higher in the normal mesenchymal cells.

	Mesenchymal tumour cells VS Epithelial tumour cells						Mesenchymal tumour cells VS Normal mesenchymal cells						Mesenchymal tumour cells VS Squamous tumour cells		
	MpBC1		MpBC2		MpBC5		MpBC1		MpBC2		MpBC6		MpBC3		
Genes	p-value	Log2FC	p-value	Log2FC	p-value	Log2FC	p-value	Log2FC	p-value	Log2FC	p-value	Log2FC	p-value	Log2FC	Total
PCOLCE	+++	1.12	+++	1.21	+++	0.32	+++	1.11	+++	1.81	+++	2.11	+++	1.40	**6**
VIM	+++	1.06	+++	1.23	0.51	-0.06	+++	1.77	+++	1.80	+++	1.51	+++	2.91	**6**
TIMP1	+++	1.10	+++	1.81	+++	0.55	+++	0.69	+++	1.43	+++	1.03	+++	2.57	**5**
AEBP1	+++	1.95	+++	1.15	0.89	-0.06	+++	2.62	+++	2.12	+++	1.05	+++	0.62	**5**
CTSK	+++	1.86	+++	1.10	0.66	-0.08	+++	2.72	+++	1.57	+++	0.98	+++	1.05	**5**
EMILIN1	+++	1.02	+++	1.09	+++	0.10	+++	1.32	+++	1.70	+++	2.25	+++	0.67	**5**
ITM2C	0.55	-0.02	+++	1.43	0.21	-0.12	+++	1.14	+++	1.98	+++	3.64	+++	1.14	**5**
PDGFRB	+++	1.37	+++	1.49	+++	0.18	+++	1.26	+++	1.79	++	0.04	+++	1.43	**5**

Table 1: Biomarkers of mesenchymal tumoural transdifferentiation (Unreliable results)
This result was obtained before noticing that there was a mistake on one of our patients' data annotations. In fact, these are the results that enabled us to detect the issue in our annotations of MpBC5 data.
In green : log2FC >= 1, In orange : 0.5 < log2FC < 1, In red : log2FC < 0.5

	Mesenchymal tumour cells VS Epithelial tumour cells						Mesenchymal tumour cells VS Normal mesenchymal cells								Mesenchymal tumour cells VS Squamous tumour cells		
	MpBC1		MpBC2		MpBC5		MpBC1		MpBC2		MpBC5		MpBC6		MpBC3		
Genes	p-value	Log2FC	p-value	Log2FC	p-value	Log2FC	p-value	Log2FC	p-value	Log2FC	p-value	Log2FC	p-value	Log2FC	p-value	Log2FC	Total
MMP14	+++	1.51	+++	1.18	+++	0.19	+++	1.92	+++	2.30	+++	1.09	+++	1.90	+++	1.91	7
PCOLCE	+++	1.12	+++	1.21	+++	0.69	+++	1.11	+++	1.81	+++	0.34	+++	2.11	+++	1.40	6
TIMP1	+++	1.10	+++	1.81	+++	1.18	+++	0.69	+++	1.42	+++	0.38	+++	1.03	+++	2.57	6
VIM	+++	1.06	+++	1.23	+++	0.51	+++	1.77	+++	1.80	+++	0.64	+++	1.51	+++	2.91	6
TNC	+++	1.01	+++	1.04	+	0.26	+++	0.99	+++	1.49	+++	1.30	+++	1.07	+++	1.13	6
AEBP1	+++	1.95	+++	1.15	+++	0.65	+++	2.62	+++	2.12	+++	0.79	+++	1.05	+++	0.62	5
CTSK	+++	1.86	+++	1.10	+++	0.66	+++	2.72	+++	1.57	+++	0.81	+++	0.98	+++	1.05	5
EMILIN1	+++	1.02	+++	1.09	+++	0.74	+++	1.32	+++	1.70	+++	0.62	+++	2.25	+++	0.67	5
CTHRC1	+++	0.92	+++	0.50	+++	0.63	+++	1.14	+++	1.76	+++	1.09	+++	1.97	+++	2.39	5
MRC2	+++	1.23	+++	1.01	+++	0.43	+++	1.69	+++	1.89	+++	0.86	+++	2.09	+++	0.75	5
NID1	+++	1.34	+++	1.05	0.21	-0.06	+++	1.41	+++	1.38	+++	0.73	+++	1.31	+++	0.61	5
PDGFRB	+++	1.37	+++	1.49	+++	0.74	+++	1.26	+++	1.79	+++	0.52	+++	0.04	+++	1.43	5
SERPINH1	+++	0.89	+++	0.99	+	-0.14	+++	1.79	+++	2.31	+++	1.08	+++	2.03	+++	2.20	5
MMP11	+++	1.17	+++	0.56	+++	0.62	+++	1.37	+++	2.28	+++	1.55	+++	0.17	+++	1.15	5
ITM2C	0.55	-0.02	+++	1.43	+++	0.27	+++	1.14	+++	1.98	+++	0.50	+++	3.64	+++	1.14	5

Table 2: Biomarkers of mesenchymal tumoural transdifferentiation

In green : log2FC >= 1, In orange : 0.5 < log2FC < 1, In red : log2FC < 0.5

MAST differential gene expression model revealed 14 potential biomarkers of the mesenchymal tumour compartment when looking only at the gene expression levels with violin plots (Fig.10). We then applied more stringent criteria to select the best biomarkers: we selected genes with a log2FC > 1 in the mesenchymal tumour cells compared to the epithelial tumoural, normal mesenchymal or squamous tumoural spots in at least 5 comparisons (table.1). This resulted in only 8 high-quality biomarkers such as PCOLCE. We can notice that patient MpBC5 has no statistically significant difference between its mesenchymal tumour compartments and the epithelial tumour ones with p-values > 0.05 for almost all the genes selected. On top of that MpBC5 is the one with log2FC < 0.5 for 7 of the 8 biomarkers identified. This has led us to question the quality of the annotations of MpBC5 data. When looking at the results of the biomarkers obtained when we corrected the annotations (table.2), we now have 15 potential biomarkers (8 of them are in common with the results obtained in table.1). The incorrect annotation of MpBC5 could've made us miss 7 potentially interesting biomarkers such as MMP14 which is by the way the best biomarker detected with p-values and log2FC that are significant (p-value < 0.001 & log2FC >= 1) in 7 out of 8 comparisons studied. Having set objective criteria instead of selecting biomarkers only based on the distributions observed on the violin plots, enabled us to avoid false positives such as gene CTH11 (Fig.10).

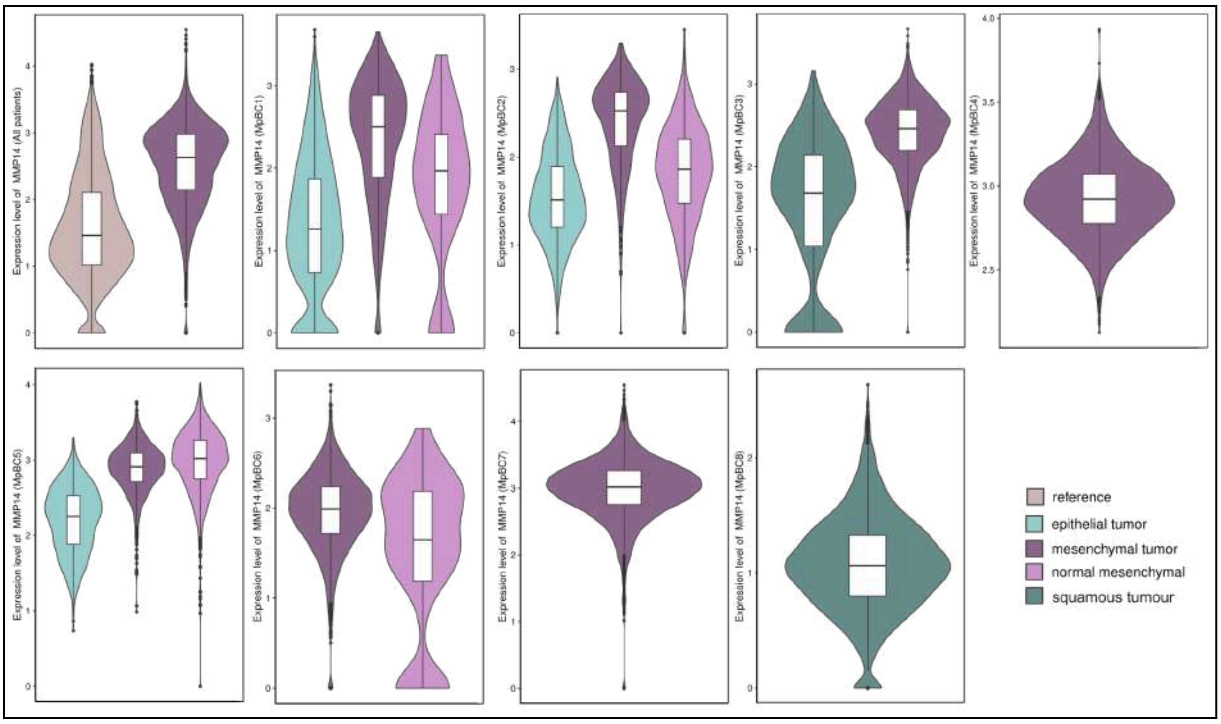


Figure 11: Example of the MMP14 gene's expression levels per patients (and per compartments)

When looking at the biomarkers identified after filtering on the per patient comparisons, better distributions of expression levels can be observed like in the example of the gene MMP14 (Fig11).

Discussions

Deconvolution clearly shows us how complex MpBC tumour structures are, it varies from one patient to the other. As we have observed, CAFs and epithelial tumour cells show spatial patterns that are mostly exclusive. The mesenchymal compartment microenvironment differs from one patient to the other, making it hard to determine a certain microenvironmental pattern that could explain why these trans-differentiations occur.

There seems to be a tendency for a more prevalent immune infiltrate in squamous cell compartments, while spindle cell and normal mesenchymal compartments display the least infiltration from immune cells. In our study, squamous cells were furthermore often located near necrotic regions. This could be in fact a result of biological mechanisms that induce this difference, such as an impact of the immune microenvironment on transdifferentiation or a pro-immune context driven by transdifferentiation, or just an artefact induced by how the slides were selected. Immune infiltration level in tumours predicts therapy response and so survival[21]. TNBC is characterised by the presence of mesenchymal tumoural cells that seems according to our results to not be infiltrated by immune cells such as T-lymphocytes which could potentially contribute to TNBC poor prognosis and survival. This will be further explored by future studies in the laboratory.

The presence of trans-differentiated mesenchymal tumour cells in the breast tissue seems cannot clearly be attributed to genomic alterations. The CNA score analysis results suggest that it is more complex than expected. In fact, most of the patients do not present major differences in CNAs between their cell compartments. However, when studying the z-scores of the CNAs, some patients do present strong outliers, suggesting that somatic genomic alterations may have occurred. They however remain very focal in most cases, possibly suggesting false positives. Only patient MpBC6 presented extended significant differences between two compartments, however both are annotated as already trans differentiated spindle cells. Overall, this suggests that genomic alterations are unlikely to be a causal explanation for the occurrence of spindle cell transdifferentiation in MpBC. Limitations in our study are definitely the size of the cohort, as well as the use of RNA-derived copy numbers rather than more straightforward DNA-based analyses. In fact the mutations in copy numbers can be a result of transcriptional differences in co-localising cis-regulatory elements, rather than genuine genomic differences.

Among GSEA results, the top one is the upregulation of the “external encapsulating structure organisation” pathway[22]. This metabolism takes place at the cellular level and more precisely outside the plasma membrane. It is responsible for the extracellular structures organisation. The Collagen metabolic process encompasses all the biological processes that involve collagen[23], which is the main component of the extracellular[24] matrix and thus very important for migration, which is highly associated with mesenchymal cells. These results are thus a confirmation that our data and their analyses were appropriate, and support the relevance of our less obvious findings involving extracellular matrix remodelling.

The up-regulation of the degradation of the extracellular matrix pathway is linked to the well documented increased motility of mesenchymal (tumour) cells, and its prominent role in metastasis. In fact, the ECM (extracellular matrix) is an important component of the microenvironment that can control cell proliferation, migration but also differentiation processes such as angiogenesis[25]. Additionally, abnormal ECM dynamics can lead to deregulated cell proliferation and invasion[25].

Among the biomarkers identified, 1 of them stands out from all since it is significantly differentially expressed with a strong fold change (log2FC >= 1) for 7 comparisons out of 8 studied: MMP14 (matrix metallopeptidase 14). MMP14 has a role in tumour progression and metastasis. Metastasis is responsible for poor prognosis. PCOLCE was found in 2022 to be correlated with cancer-associated fibroblasts infiltration[26]. This piece of information consolidates our intuition that the TNBC tumour could be influenced by the microenvironment. Interestingly, TIMP1 (tissue inhibitor of metalloproteinases 1) overexpression has been proven in 2023 to be linked with chemoresistance in mesenchymal TNBC[27]. VIM(Vimentin) has been found to be related to TNBC cancer cells' resistance to chemotherapy and is presented as a possible target in drug resistance and resurgence of the disease [28]. TNC (Tenascin-C) is a gene coding for an extracellular matrix glycoprotein that is involved in cell proliferation. Its overexpression has also been linked to poor prognosis. Interestingly, the loss-of-function of TNC showed significant decrease of cell proliferation and even reversed the mesenchymal phenotypes to epithelial ones in breast cancer tumour cells[29]. AEBP1 (adipocyte enhancer binding protein 1) is also described as involved in cell proliferation in breast cancer [30]. Paradoxically, we observe a significant overexpression of EMILIN1 (elastin microfibril interfacer 1) in our TNBC samples when in literature it is said that its downregulation is common with breast cancer[31]. This gene is involved in extracellular matrix organisation, collagen metabolic process and defence response. CTSK (cathepsin K) has been shown to be “strongly expressed in human breast cancers with primary or developing bone metastases”[32]. CTHRC1 (collagen triple helix repeat containing-1) is produced by CAFs and has a role in cancer cells invasiveness and EMT (epithelial mesenchymal transition). CAFs seem to be responsible for aggressive behaviour of tumour via CTHRC1/wnt/Beta-catenin signalling pathway[33]. MRC2(Mannose receptor C type 2) plays a role in extracellular matrix remodelling. This gene’s expression is said to possibly be involved in tumorigenesis and metastasis[34]. NID1 (Nidogen 1) “may play a role in cell interactions with the extracellular matrix”[35]. A study suggests that PDGFRB (Platelet-derived growth factor receptor Beta) could be a “a unique feature of cancer cells that possess stem cell characteristics and/or that have undergone EMT, two features associated with chemoresistance and aggressiveness, thus encouraging therapeutic options in targeting this receptor” [36]. SERPINH1 (Heat shock protein 47) could be involved in the pathway of collagen according to the Uniprot database. MMP11 (Matrix metalloproteinase 11) is a gene involved in breast cancer tumour progression[24]. Lastly, ITM2C (Integral membrane protein 2C) may play a role in TNF-induced cell death[37].

Spatially resolved transcriptomics enables us to understand diseases more globally from the study of the microenvironment, to the genomic expression profiles of cells. The integration of scRNAseq data along with spatial data enables it to go from mixed sample, bulk resolution to a single cell one. Nonetheless, the results produced present some limitations that are inherent to the transcriptomic nature of the data. In fact, this technology suffers from its average power of resolution. Similarly to UMI-based single cell approaches, SRT technology isn’t precise enough to detect lowly expressed genes and so there is a risk of drop-outs. Lack of precision could also be explained by spot swapping: regardless of the UMIs that are spot-specific in order to measure with precision the RNA counts, there is in reality possible bleeding of RNA from close spots. There is a package SpotClean that is supposed to provide “more accurate estimates of gene-specific UMI counts” that unfortunately couldn’t be used on our data. The use of tools such as Spotclean before undergoing further data analysis could improve the reliability of the results. The authors report that “in multiple studies of cancer, SpotClean improves tumour versus normal tissue delineation and improves tumour burden estimation thus increasing the potential for clinical and diagnostic applications of spatial transcriptomics technologies”.

The annotation error that occurred for the MpBC5 sample showed in multiple analyses how it can significantly change the results obtained and so the direction of our hypothesis and interpretations. The annotation of spots being a daunting, human-dependent task, it is prompt to errors. One solution could’ve been to annotate the spots based on the results of the deconvolution by assigning the annotation of the cell type that is the most present in a certain spot. However this idea presents limitations. In fact in order to annotate the spots, several factors have to be taken into account such as the morphology of the cells and their colouration. The deconvolution is only based on scRNAseq data. Moreover, the breast cancer atlas isn’t detailed enough. In fact, there aren’t mesenchymal tumoural cells in the breast cancer atlas. Additionally, TNBC is a cancer that isn’t well understood yet which explains why the breast cancer atlas doesn’t include the type of cells that we can find in TNBC tumours. In the long term vision, and with AI being more and more efficient in several medical fields we can think of a possible use of it for cell annotations. Additionally, another limitation of our findings is our inability to assess the biomarkers' findings reproducibility on a second validation patient cohort. Moreover, biomarkers have to undergo clinical validation before being included in routine use. Biomarker candidates' robustness depends on numerous and meticulous analyses that should be successfully passed. This lack of generalizability makes our findings not yet fully reliable before further validation. Biomarkers robustness assessment is a crucial stake because it can impact clinical care, for better or for worse. An article explains the importance of performing high levels scrunity when it comes to biomarkers selection, by giving an example of a canadian study: “it was discovered upon retesting that 40% of women originally diagnosed with oestrogen receptor negative breast cancer actually had oestrogen receptor positive tumours, and thus were deprived of a potentially life‑saving treatment”[38].

Conclusion

The aim of the present thesis was to shed light on the transdifferentiation mechanism of metaplastic breast tumours. The most significant finding that emerges from this study is that Triple negative breast cancer is a very complex disease, where microenvironment seems to play an important role, more than genomics, in influencing the change in phenotype of tumoural cells. Mesenchymal tumour compartments seem to be, as opposed to squamous tumoural ones, barely infiltrated by immune cells which could explain the very aggressive trait of these cells. Additionally, some potential biomarkers as well as key pathways have been identified. Extracellular matrix organisation related pathways up-regulation could explain the likeliness of TNBC tumour cells to invade other distant organs such as the lungs or brain.

Despite the limitations that are induced by the SRT methodology or by the limited amount of data that is responsible of a lack of reliability of our findings, we can at least successfully came up with new suggestions by finding the adapted bioinformatic tools such RCTD powerful deconvolution R package, MAST or Spotclean for future use, that can hereafter be used by researchers in the study.

Further studies need to be carried out taking into consideration the challenges faced as well as the limitations of our current findings. In the future, the team is planning on undertaking microdissection, in order to use more precise DNA-based methods to better study the evolutionary history of these MpBC samples. They are also planning on increasing the cohort size in order to increase accuracy and statistical power..

Bibliography

1. Monaco, M. L., Idris, O. A. & Essani, K. Triple-Negative Breast Cancer: Basic Biology and Immuno-Oncolytic Viruses. *Cancers* **15**, 2393 (2023).
2. Siegel, R. L., Miller, K. D., Wagle, N. S. & Jemal, A. Cancer statistics, 2023. *CA. Cancer J. Clin.* **73**, 17–48 (2023).
3. Coutant, A. *et al.* Spatial Transcriptomics Reveal Pitfalls and Opportunities for the Detection of Rare High-Plasticity Breast Cancer Subtypes. *Lab. Investig. J. Tech. Methods Pathol.* **103**, 100258 (2023).
4. Triple-negative Breast Cancer | Details, Diagnosis, and Signs. https://www.cancer.org/cancer/types/breast-cancer/about/types-of-breast-cancer/triple-negative.html.
5. What Causes Breast Cancer? https://www.cancer.org/cancer/types/breast-cancer/about/how-does-breast-cancer-form.html.
6. 10x_LIT000128_Product_Sheet_Spatial_Biology_Without_Limits_Letter_digital.pdf.
7. Williams, C. G., Lee, H. J., Asatsuma, T., Vento-Tormo, R. & Haque, A. An introduction to spatial transcriptomics for biomedical research. *Genome Med.* **14**, 68 (2022).
8. Coleman, K., Hu, J., Schroeder, A., Lee, E. B. & Li, M. SpaDecon. https://doi.org/10.5281/zenodo.7735251 (2023).
9. Coleman, K., Hu, J., Schroeder, A., Lee, E. B. & Li, M. SpaDecon: cell-type deconvolution in spatial transcriptomics with semi-supervised learning. *Commun. Biol.* **6**, 1–13 (2023).
10. Wu, S. Z. *et al.* A single-cell and spatially resolved atlas of human breast cancers. *Nat. Genet.* **53**, 1334–1347 (2021).
11. Tools for Single Cell Genomics. https://satijalab.org/seurat/.
12. Cable, D. dmcable/spacexr. (2024).
13. CharleneZ95/infercnvPlus: Enhanced ‘infercnv’ package version 0.1.0 from GitHub. https://rdrr.io/github/CharleneZ95/infercnvPlus/.
14. Cytoband Hg38 - 2022.
15. FindMarkers function - RDocumentation. https://www.rdocumentation.org/packages/Seurat/versions/1.3/topics/FindMarkers.
16. EnhancedVolcano: publication-ready volcano plots with enhanced colouring and labeling. https://bioconductor.org/packages/release/bioc/vignettes/EnhancedVolcano/inst/doc/EnhancedVolcano.html.
17. Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 278 (2015).
18. How to do Gene Set Enrichment Analysis (GSEA) in R. *Bioinformatics Breakdown* https://bioinformaticsbreakdown.com/how-to-gsea/ (2019).
19. fgsea. *Bioconductor* http://bioconductor.org/packages/fgsea/.
20. create.RCTD: Creates an ‘RCTD’ object from a scRNA-seq reference... in dmcable/RCTD: SpatialeXpressionR: Cell type identification and cell type-specific differential expression in spatial transcriptomics. https://rdrr.io/github/dmcable/RCTD/man/create.RCTD.html.
21. Zou, R. *et al.* Characteristics of Infiltrating Immune Cells and a Predictive Immune Model for Cervical Cancer. *J. Cancer* **12**, 3501–3514 (2021).
22. GOBP_EXTERNAL_ENCAPSULATING_STRUCTURE_ORGANIZATION. https://www.gsea-msigdb.org/gsea/msigdb/cards/GOBP_EXTERNAL_ENCAPSULATING_STRUCTURE_ORGANIZATION.

23. GOBP_COLLAGEN_METABOLIC_PROCESS. https://www.gsea-msigdb.org/gsea/msigdb/cards/GOBP_COLLAGEN_METABOLIC_PROCESS.
24. Ling, B. *et al.* A novel immunotherapy targeting MMP-14 limits hypoxia, immune suppression and metastasis in triple-negative breast cancer models. *Oncotarget* **8**, 58372–58385 (2017).
25. Reactome | Extracellular matrix organization. https://www.reactome.org/content/detail/R-HSA-1474244.
26. Gao, H. & Li, Q. A pan-cancer analysis of the oncogenic role of procollagen C-endopeptidase enhancer (PCOLCE) in human. *Medicine (Baltimore)* **101**, e32444 (2022).
27. Agnello, L. *et al.* Tissue Inhibitor of Metalloproteinases-1 Overexpression Mediates Chemoresistance in Triple-Negative Breast Cancer Cells. *Cells* **12**, 1809 (2023).
28. Winter, M. *et al.* Vimentin Promotes the Aggressiveness of Triple Negative Breast Cancer Cells Surviving Chemotherapeutic Treatment. *Cells* **10**, 1504 (2021).
29. Wawrzyniak, D. *et al.* Down-regulation of tenascin-C inhibits breast cancer cells development by cell growth, migration, and adhesion impairment. *PLoS ONE* **15**, e0237889 (2020).
30. Li, J. *et al.* AEBP1 Contributes to Breast Cancer Progression by Facilitating Cell Proliferation, Migration, Invasion, and Blocking Apoptosis. *Discov. Med.* **35**, 45–56 (2023).
31. Favero, A. *et al.* Loss of the extracellular matrix glycoprotein EMILIN1 accelerates Δ16HER2-driven breast cancer initiation in mice. *Npj Breast Cancer* **10**, 1–13 (2024).
32. Qian, D. *et al.* Cathepsin K: A Versatile Potential Biomarker and Therapeutic Target for Various Cancers. *Curr. Oncol.* **29**, 5963–5987 (2022).
33. Li, H., Liu, W., Zhang, X. & Wang, Y. Cancer‑associated fibroblast‑secreted collagen triple helix repeat containing‑1 promotes breast cancer cell migration, invasiveness and epithelial‑mesenchymal transition by activating<br />the Wnt/β‑catenin pathway. *Oncol. Lett.* **22**, 1–10 (2021).
34. MRC2 mannose receptor C-type 2 [Homo sapiens (human)] - Gene - NCBI. https://www.ncbi.nlm.nih.gov/gene/9902#summary.
35. NID1 nidogen 1 [Homo sapiens (human)] - Gene - NCBI. https://www.ncbi.nlm.nih.gov/gene/4811.
36. Camorani, S. *et al.* Targeted imaging and inhibition of triple-negative breast cancer metastases by a PDGFRβ aptamer. *Theranostics* **8**, 5178–5199 (2018).
37. ITM2C - Integral membrane protein 2C - Homo sapiens (Human) | UniProtKB | UniProt. https://www.uniprot.org/uniprotkb/Q9NQX7/entry.
38. Henry, N. L. & Hayes, D. F. Cancer biomarkers. *Mol. Oncol.* **6**, 140–146 (2012).

Abstract

Key words : Metaplastic breast cancer, transdifferentiation, biomarkers, pathways, immune infiltration

Triple negative breast cancer (TNBC) is both a rare and aggressive subtype, representing between 10 and 20% of all breast cancer cases, and is unfortunately still not well characterised yet. To date, no biomarkers driving the very specific carcinogenesis mechanism of the TNBC tumours have been identified making it ineligible to targeted therapies and so the most lethal breast cancer subtype. This study is aimed to provide new insights on the transdifferentiation of metaplastic tumours, with a specific focus on mesenchymal spindle cell tumour cells. Metaplastic breast carcinoma (MpBC) is characterised by an abnormal transdifferentiation of epithelial cells to a different phenotype, whose presence is not expected in the tissue. The primary goal was to analyse the genomic, transcriptomic as well as the microenvironmental differences between the different trans-differentiated compartments. Ultimately, this research seeks to determine the genes and pathways involved in the transdifferentiation mechanism and potential specific biomarkers of these tumours. In order to unravel the mechanisms that drive metaplastic tumours, we studied spatial transcriptomics data (10x Genomics Visium), collected from 8 MpBC patients. Spatially resolved transcriptomics approach captures both histological and transcriptional information.

Résumé

Mots-clés : Cancer du sein métaplasique, transdifferentiation, biomarqueurs, infiltration immunitaire

Le cancer du sein triple négatif est un sous-type à la fois rare et agressif, qui représente entre 10 et 20 % de tous les cas de cancer du sein, et qui n'est malheureusement pas encore bien caractérisé. À ce jour, aucun biomarqueur du mécanisme de carcinogenèse très spécifique des tumeurs TNBC n'a été identifié, ce qui les rend inéligibles aux thérapies ciblées et en fait le sous-type de cancer du sein le plus létal. Cette étude vise à fournir de nouvelles informations sur la transdifférenciation des tumeurs métaplasiques, en mettant l'accent sur les cellules mésenchymateuses des tumeurs à cellules fusiformes. Le carcinome métaplasique du sein se caractérise par une transdifférenciation anormale des cellules épithéliales en un phénotype différent, dont la présence n'est pas attendue dans le tissu. L'objectif principal était d'analyser les différences génomiques, transcriptomiques et micro environnementales entre les différents compartiments trans différenciés. En fin de compte, cette recherche vise à déterminer les gènes et les voies impliqués dans le mécanisme de transdifférenciation et les biomarqueurs spécifiques potentiels de ces tumeurs. Afin d'élucider les mécanismes à l'origine des tumeurs métaplasiques, nous avons étudié les données transcriptomiques spatiales (10x Genomics Visium), recueillies auprès de 8 patientes atteintes de cancers métaplasiques. L'approche transcriptomique spatiale capture à la fois les informations histologiques et transcriptionnelles.