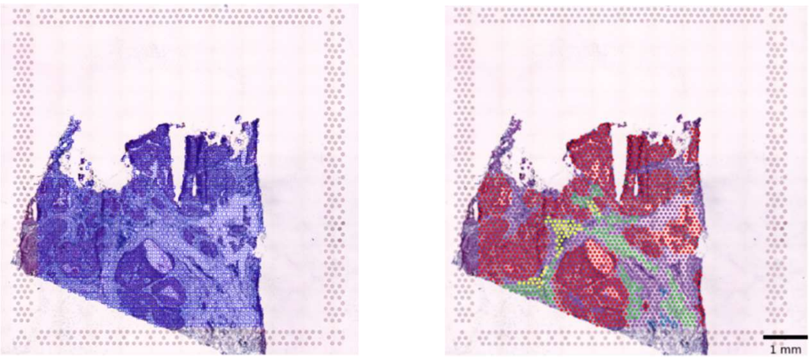
Internship report

Academic year: 2023-2024

Corentin Meyvaert

Integration of histopathological image data and expression data from Spatial Transcriptomics to identify tissue types and cancer in breast tissue sections



Supervised by Pr. Carsten Daub

Karolinska Institutet, NEO, Blickagangen, Huddinge, Sweden

Acknowledgements

I would like to express my gratitude to everyone who welcomed and assisted me during my internship. Firstly, I would like to thank Carsten Daub, who guided and mentored me throughout my entire internship. Thank you for the trust and confidence you placed in me. I would also like to thank all the member of the group, who were very friendly and who took the time to introduce me to their project and provide me with valuable feedback. I would also like to thank all the people from the ST group, for all these very informative and interesting meetings. Thank you for giving me the opportunity to present several times and thank you for all your detailed advice which helped me improve my project. Finally, I am very thankful for this time spent in Sweden, and I appreciate all the wonderful Fika moments.

Table of Contents

Résumé :

Le diagnostic du cancer du sein est généralement réalisé par un pathologiste expert examinant la morphologie d'une section de tissu mammaire. Au début du XXIe siècle, le développement des techniques de transcriptomique a permis une caractérisation approfondie du cancer du sein en analysant l'expression des gènes dans les tissus cancéreux. Cependant, identifier l'hétérogénéité intra- et inter-tumorale dans le cancer du sein reste un défi, et caractériser cette hétérogénéité moléculaire est essentiel pour améliorer le diagnostic, définir des biomarqueurs pronostiques et concevoir des stratégies thérapeutiques. La transcriptomique spatiale est une technique innovante utilisée pour analyser les profils d'expression génique à différents endroits au sein des sections de tissu. Dans ce rapport, les données d'images histopathologiques et les données d'expression issues de la transcriptomique spatiale sont combinées et exploitées à l'aide de machine learning pour identifier les types de tissus dans des sections de cancer du sein. Les résultats montrent que les données de transcription résolues spatialement peuvent identifier les tissus cancéreux, mais sont moins efficaces pour distinguer les différents types de tissus non tumoraux. La caractérisation des marqueurs d'expression génique indique que l'identification des types de tissus à l'aide des données de transcriptomique spatiale peut dépendre de motifs d'expression génique cohérents. Les évaluations des performances diagnostiques utilisant des images histopathologiques ont démontré que le cancer et divers types de tissus peuvent être identifiés, en particulier lorsque des modèles de deep learning sont entraînés sur des images histopathologiques. Cependant, l'intégration des données d'expression résolues spatialement avec les données d'images histopathologiques n'a pas montré d'amélioration significative dans la détection des tissus. Des modèles d'apprentissage automatique non supervisés peuvent distinguer les types de tissus non cancéreux des types de tissus cancéreux. Dans l'ensemble, la détection du cancer utilisant la transcriptomique spatiale et le machine learning semble prometteuse pour l'avenir du diagnostic du cancer.

Mots-clés : cancer du sein, image de section de tissue, transcriptomique spatiale, intégration de données, machine learning

Abstract:

Breast cancer diagnosis is commonly performed by an expert pathologist examining the morphology of a breast tissue section. In the early 21st century, the development of transcriptomics techniques enabled an in-depth characterization of breast cancer by analyzing the gene expression of cancerous tissues. However, identifying both intra- and inter-tumor heterogeneity in breast cancer remains challenging, and characterizing this molecular heterogeneity is crucial for improving diagnosis, defining prognostic biomarkers, and designing therapeutic strategies. Spatial transcriptomics is an innovative technique used to analyze gene expression profiles at different locations within tissue sections. In this report, histopathological image data and expression data from spatial transcriptomics are combined and leveraged through machine learning to identify tissue types within breast cancer sections. The results show that spatially resolved transcription data can identify cancerous tissues but are less effective in distinguishing different non-tumoral tissue types. The characterization of gene expression markers indicates that identifying tissue types using spatial transcriptomics data may depend on coherent gene expression patterns. Performance evaluations of diagnostics using histopathological images have demonstrated that cancer and various tissue types can be identified, particularly when deep learning models are trained on histopathological images. However, integrating spatially resolved expression data with histopathological image data has not shown a significant improvement in tissue detection. Unsupervised machine learning models may distinguish between non-cancerous and cancerous tissue types. Overall, cancer detection using spatial transcriptomics and machine learning appears promising for the future of cancer diagnosis.

Keywords: breast cancer, tissue section image, spatial transcriptomics, data integration, machine learning

Abbreviation

MRI: Magnetic Resonance Imaging

ST: Spatial Transcriptomics

ML: Machine Learning

AI: Artificial Intelligence

CBB: Center of Bioinformatics and Biostatistics

ER+: Estrogen Receptor positive

BB: Benign Breast

IC: Immune Cells

H&E: Hematoxylin and eosin

RT: Reverse Transcription

NGS: Next Generation Sequencing

UMI: Unique Molecular Identifier

DEA: Differential Expression Analysis

GO: Gene Ontology

FCNN: Fully Convolutional Neural Network

CNN: Convolutional Neural Network

RF: Random Forest

ARI: Adjusted Rand Index

DCIS: Ductal Carcinoma In Situ

ILC : Invasive Lobular Carcinoma

IDC : Invasive Ductal Carcinoma

TD: Transcriptomics Data

SID: Stardist Image Data

KID : KimiaNet Image Data

I. Introduction

1. Context of the study

a. Breast Cancer

Breast cancer is a disease in which abnormal breast cells grow out of control and form tumors. If left unchecked, the tumors can spread throughout the body and become fatal. This disease caused 670 000 deaths globally in 2022. Roughly half of all breast cancers occur in women with no specific risk factors other than sex and age. Breast cancer was the most common cancer in women in 157 countries out of 185 in 2022.[1]

The disease is caused by the apparition of cancer cells inside the milk ducts and/or the milk-producing lobules of the breast. The earliest form is not life-threatening and can be detected in early stages. Cancer cells can spread into nearby breast tissue (invasion). This creates tumors that cause lumps or thickening.[1]

If a breast cancer is detected at an early stage, there is 99% chance of survival. Otherwise, if the cancer is identified at a late stage, there is 26% chance of survival.[2] Early diagnosis of breast cancer is therefore a significant challenge to increase the survival rate of the patients.

There are several diagnosis methods:

- Clinical examination by a healthcare provider, discussion of the symptoms and the medical history.[3]
- Imaging tests: these tests help visualize breast tissue and identify any abnormalities (e.g. Mammogram, Breast MRI, Breast Ultrasound).[4]
- Biopsy: if an abnormality is detected, a sample of breast tissue is removed for testing.[3]
- Genomic Lab Tests: these tests analyze genetic material to understand the specific characteristics of the cancer.[5]

For the biopsy, an expert (pathologist) needs to analyze the sample and to find if there are some cancer cells in the tissue.

However, it is a challenging task to recognize tens of thousands of histopathological images of liver cancer by naked eye, which poses numerous challenges to inexperienced clinicians.

Identification of both intra- and inter-tumor heterogeneity in breast cancer poses a significant challenge due to its genomic evolution that occurs during breast cancer progression. In depth characterization of the molecular heterogeneity is important to improve diagnosis, define prognostic biomarkers and for designing therapeutic strategies.[6]

b. Spatial transcriptomics

Bulk transcriptomics is the study of all the RNA present at a certain time in a cell or in a tissue. It allows to know which genes are expressed, at what levels, and under what conditions. It is useful for understanding cellular mechanisms, developmental processes, responses to environmental stimuli, or diseases. However, bulk transcriptomics come from homogenized biopsies, which gives an average transcriptome of a tissue. Therefore, there is a loss of the spatial information, which can be important to study complex tissues consisting of various cell types or for cell types that are rare in a tissue.[7]

Spatial Transcriptomics (ST) is a transcriptomics method that permits to maintain spatial information of the tissue section. It improves the understanding of tissue functionality and corresponding pathological changes. Moreover, it is efficient to study the intratumor heterogeneity that fosters tumor evolution, and it also allows to compare different areas within a tissue section.[7]

c. Machine learning

Machine Learning (ML) is a subfield of Artificial Intelligence (AI), which is broadly defined as the capability of a machine to imitate intelligent human behavior.[8] ML is the field of study that gives computers the ability to learn without explicitly being programmed. It is now broadly use to solve many real-world problems, and especially in the Health and Biology domains, for its capabilities to process huge amounts of data.[9] For example, ML can be used to analyze gene expression, to predict protein structure and to identify patterns in patient data to predict disease risk. Multimodal ML is the name given to a ML algorithm when various data types are combined with multiple intelligence processing algorithms to achieve higher performances. Multimodal ML often outperforms single modal ML in many real-world problems.[10] ML's ability to learn from data continuously enhances accuracy, paving the way for more effective, data-driven decisions in both research and clinical settings.

d. New technologies to improve breast cancer prognosis

Breast cancer diagnosis remains a critical challenge, as early detection is key to improving patient outcomes. To enhance accuracy and detect the disease in its early stages, innovative approaches are urgently needed. Emerging technologies like ST offer data by mapping gene expression to tissue architecture, providing spatially resolved insights that were previously unattainable. Machine learning can play a pivotal role in interpreting these complex ST datasets, enabling automated and more precise diagnoses. By leveraging ML, we can significantly advance the diagnostic process, potentially transforming breast cancer detection and ultimately improving patient prognosis.

2. Goals

Overall goal: Can the integration of histopathological image data and expression data from ST be leveraged by machine learning to identify tissues within breast cancer sections?

Specific goals:

1. Identify tissue types in histopathology slides using spatially resolved expression data.
2. Characterize tissue types with gene expression markers.
3. Identify tissue types in histopathology slides using histopathological image data.
4. Enhance tissue type identification through integration of spatially resolved expression data and histopathological image data.
5. Recognize tissue types with unsupervised machine learning using spatial transcriptomics data and histopathological images.

II. Materials and methods

1. Dataset

a. Preview of the dataset

The dataset I used for my work is a ST dataset composed of 11 breast cancer patients who were diagnosed with ductal carcinoma in situ (DCIS), invasive lobular carcinoma (ILC) or invasive ductal carcinoma (IDC). 10 out of these 11 patients are Estrogen Receptor positive (ER +). From these patients, 48 slides were built, with some slides from tumor and some from tissue near tumor. And for each of these slides, image of the tissue and spatial transcriptomics data are available. For this research project, the focus was mainly on patient 3 (4 slides - 2 non tumoral, 2 tumoral - ER +, IDC). For some test, slides from other patients were also used.

Microscope images are 8.37 mm width and 8,11 mm height. They are in high resolution: 48640 and 47104 pixels, respectively for width and height (which correspond to approximately 148 DPI). Images are used as an anatomical map for expression data.

Spatial transcriptomics slides are composed of spots, where the total expression of cells is measured for each spot separately. The spots diameters are 65 µm, which is approximately equal to 40 cells per spots. 36601 genes expression is measured for each spot, and there is approximately 2000 of spot on each slide.

For each spot, an expert pathologist has annotated the tissue type by looking at the anatomical organization in the image at the position of the spot. The annotation provided are the following: Benign breast (BB), Stroma, Fat, Tumor, Immune cells (IC), Mixed and Out. Mixed is the annotation given when a spot contains different cell types (non-homogeneous cell type). Out stands for spot where the image has a bad quality and where a cell type cannot be determined (for example because of a bad staining). Out spots were removed from the analysis.

b. Building the dataset

Samples preparation:

For each patient, tissue samples were taken with a mastectomy surgery. Then we got tissue section of each tissue sample, which were embedded and covered using OCT. Then they were sectioned on slides of 6,5 mm x 6,5 mm (ST slide) by using a cryostat. Each section has been stained using Hematoxylin and eosin (H&E) coloration. Finally, each section has been imaged.

Spatial transcriptomics experimentation:

Visium Spatial Gene Expression Slide were used to perform the ST experimentation. These slides contain 5000 barcoded spots which can capture different area of the tissue and their expression. Each spot contains probes with a spatial barcode, a Unique Molecular Identifier (UMI), and a Poly(dt) nucleotide sequence. The tissue section is placed on the slide and permeabilized, RNA is released and binds to the capture probes, allowing for capture of the gene expression information, and then a reverse transcription (RT) is carried out to get barcoded cDNA. Finally, the probes are cleaved from the slides and a Next Generation Sequencing (NGS) is carried out. The cDNA fragments obtained from the RT are sequenced by Illumina's NGS with a depth 500X in paired end. Read one contains the spatial barcode and the UMI. Read two contains the transcript sequence information. Illumina's outputs are files in FASTQ format; they contain the sequence and the quality of each read.

2. Data pre-processing

The ST outputs were subsequently processed using the Space Ranger Pipeline. This pipeline generates several outputs by utilizing the FASTQ files along with images of each section. These outputs contain various information, including the expression count matrix for each spot or the position on the image of each barcode.

To achieve this, Space Ranger is structured into different pipelines, two of which were utilized:

1. **Spaceranger count**: This pipeline takes the slide images and FASTQ files, aligns the reads to the human genome (using STAR), and counts the barcodes/UMIs. It employs spatial barcodes to generate count matrices, where a count matrix reflects the number of unique gene observations in each spot barcode. The count of each mRNA in a spot is divided by the number of its UMIs, which removes PCR duplicates. In this matrix, transcripts are represented by rows, while each barcode is a column. Moreover, the coordinates of the spots (positions on the slide) are determined from the tissue image, and the file tissue_positions_list.csv is generated.
2. **Spaceranger aggr**: This pipeline takes the outputs from Spaceranger count and aggregates them, normalizing the data to the same sequencing depth (gene count in a spot/average spot count), and then recalculates the count matrices. After this step, two files are generated for each sample: the "filtered_feature_bc_matrix.h5" and an image of the tissue staining aligned with the spot information.

The expression is measured for 36601 genes in our dataset, which correspond to the number of genes found during the mapping with the human genome.

The outputs from Spaceranger count were used for all transcriptomics data analyses, except for the ConGI tool, where the outputs from Spaceranger aggr were utilized.

Transcription data from Spaceranger count's outputs are normalized, Log1p transformed, batch corrected and only the 350 most variables genes are kept.

This is a per-spot normalization, which means that expression of each spot was normalize separately. For each spot, it adjusts the gene values so that the sum of the gene values in that spot equals a fixed value (by default, this value is 1e4). This helps make gene measurements comparable between different spots, despite potential differences in the number of RNA molecules captured by each spot.

The Log1p is particularly useful for mitigating the effect of outliers. Moreover, by adding 1 before taking the logarithm, we avoid the problem of the logarithm of zero, which is undefined.

The 350 most variables genes are kept to identify genes whose expression variability is higher than would be expected by chance, which may indicate important biological regulation. The variable genes were calculated for each batch separately (batch correction).

3. Differential Expression Analysis and Gene Ontology

The Differential Expression Analysis (DEA) was performed on patient 3 transcriptomics data (4 slides), with the scanpy python's module[11] and with FDR adjusted P-value < 0.05, and

Log2 fold-change > 1 or Log2 fold-change < -1. The Gene Ontology (GO) analysis was performed on the significant genes found in the DEA for each tissue type with the DAVID[12] tools. Down-regulated genes and up-regulated genes have been processed separately.

4. Tiling the image

Full resolution images were subdivided in tiles of 65 x 65 µm. Tile positions correspond to ST spots, which allow to combine image data with gene expression data. Furthermore, pathologists have annotated each tile with its respective tissue type, allowing supervised ML on all annotated tiles.

The **Tissue_positions_list** from the **Spaceranger count** pipeline, which contained the pixel position of each ST spot, was used to create tiles centered on the ST spots. The pipeline also creates a **Scalefactor** file, which allows to get the diameter in units of pixel of the ST spots. The tiles' size was chosen to avoid overlapping of tiles (each ST spot has a diameter of 65 µm).

5. Segmentation and extraction of features of tiles

Morphological differences between tissue types can be used for identifying spatial domain. For example, malignant cells have irregular border, variation in shape, a large and irregular nuclei while benign cells have a well-defined border, an uniform shape and an uniform nuclei.[13] Stardist algorithm and a Fully Convolutional Neural Network (FCNN) have been used to segment tiles and to extract features from them.

a. Stardist method

Stardist is a deep learning algorithm, which is using a Convolutional Neural Network (CNN) to predict object center probabilities and boundary distances. The algorithm is then capable of reconstructing object shapes using the distance predictions. Because it can recognize round object, it is particularly useful to analyze cell nuclei in a tissue. Here are some examples of features that Stardist can extract from round object in an image: area, perimeter, eccentricity, mean intensity, solidity.[14] For the analysis, all of the 27 features were extracted to perform some ML.

b. KimiaNet method

KimiaNet is a FCNN pre-trained on more than 240,000 histopathological image patches. Images used are from The Cancer Genome Atlas and are labelled with pathologist's annotations. KimiaNet employs the topology of the DenseNet with four dense blocks, fine-tuned and trained with histopathology images in different configurations. The advantage of

KimiaNet compared to other pre-trained CNN is that it has been pre-trained on image from the biological domain and more precisely on cancer images.[15] For the analysis, 1024 features were extracted to perform some ML.

6. Machine learning

Machine learning technique were used to learn expression signatures (train a model) for different tissue in breast cancer slides. The regions were identified in two ways, by supervised expert annotation of the images and independently in an unsupervised way by Kmeans clustering. The identified expression signatures were used to characterize ST spots in ST experiments (testing of the model) which were withheld during model training.

Various machine learning methods, such as Random Forest (RF), Stacking, and CNN were used. Default values were used for the hyper-parameters of the RF, and a selection of the best parameters after some trials has been taken for Stacking and CNN methods. The result of one test consists of associations of ST spots to classes. Classification performances were evaluated by f1-score and accuracy metrics. The f1-score is calculated based on true positive and true negative results. The f1-score ranges from 1 for perfect prediction to 0. Clustering performance was evaluated by Adjusted Rand Index (ARI). The ARI scores range from 1.0 for perfect clustering to zero for random clustering, with scores below zero indicating worse clustering than what would be expected by chance.

For multimodal ML, a scaling has been done on the data, using the sklearn MinMaxScale function.

7. ConGI

ConGI is a multimodal AI tool that can integrate ST expression data and histopathological image data, by making image tiles for each ST spot. This tool was used to perform classification and clustering on the slide 3C. Here are some interesting features that ConGI has: it use PCA to reduce the dimension of gene expression to 300, it normalizes the expression data, it uses two independent encoders (a DenseNet CNN for image and an Multi-Layer Perceptron for gene expression) to learn the low-dimensional representations, it uses data augmentation algorithms, and cosine similarity between low dimensional vectors are calculated with contrastive loss (including gene expression to gene expression, image to image, and image to gene expression). At the end, it pulls the pairwise data within and between modalities together and contrasts the unmatching pairs apart.

8. Computational material

All the analysis were performed on the CBB server (Center of Bioinformatics and Biostatistics, Karolinska Institutet) with following setup:

- 2 * AMD EPYC 7302 Rome 16-Core 3.00GHz
- 2 * Nvidia RTX 6000 GPU 24GB (cuda 11.1)
- 2 TB RAM.

Majority of the code was implemented in python. The R package mclust was also used.

The code is available in the following GitLab page:

https://gitlab.com/daub-lab/corentin_internship

9. General diagram of the project

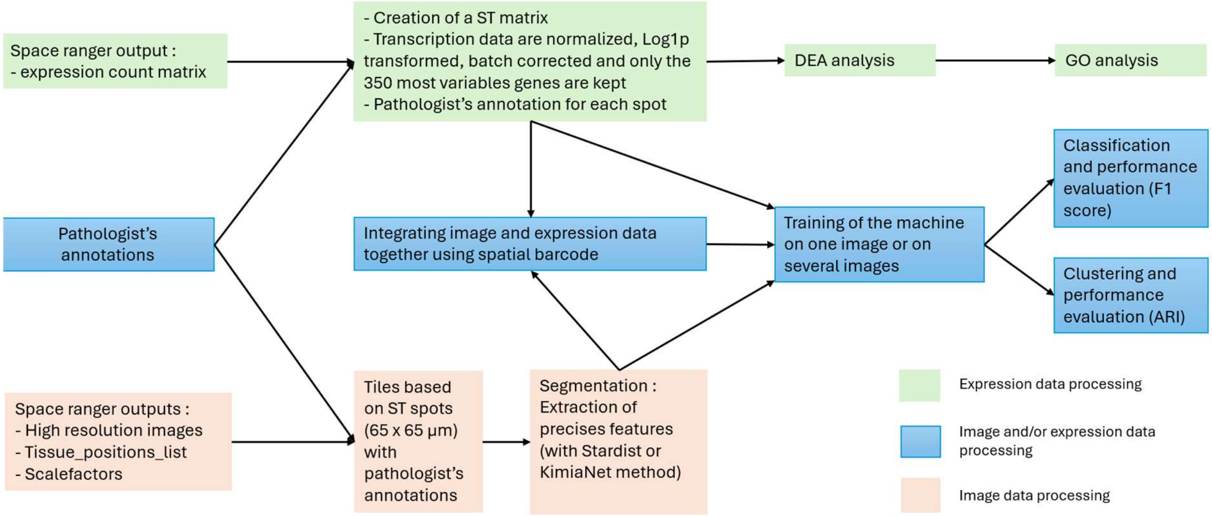

Space ranger output :
- expression count matrix
- Creation of a ST matrix
- Transcription data are normalized, Log1p transformed, batch corrected and only the 350 most variables genes are kept
- Pathologist's annotation for each spot
DEA analysis
GO analysis
Pathologist's annotations
Integrating image and expression data together using spatial barcode
Training of the machine on one image or on several images
Classification and performance evaluation (F1 score)
Clustering and performance evaluation (ARI)
Space ranger outputs :
- High resolution images
- Tissue_positions_list
- Scalefactors
Tiles based on ST spots (65 x 65 µm) with pathologist's annotations
Segmentation : Extraction of precises features (with Stardist or KimiaNet method)
Expression data processing
Image and/or expression data processing
Image data processing

Figure 1 General diagram for identifying spatial domains using image and gene expression data. Image data (in orange) and expression data (in green) were processed separately or together (in blue). The outputs from the Space Ranger count pipeline and the pathologists's annotations were used as input data. Image and gene expression data were pre-processed separately. Differential Expression Analysis (DEA) and Gene Ontology (GO) analysis were performed on the gene expression data. Various machine learning methods, such as Random Forest (RF), Stacking, and Convolutional Neural Networks (CNN), were used. Classification and clustering performance were evaluated using the F1-score and Adjusted Rand Index (ARI), respectively.

III. Results

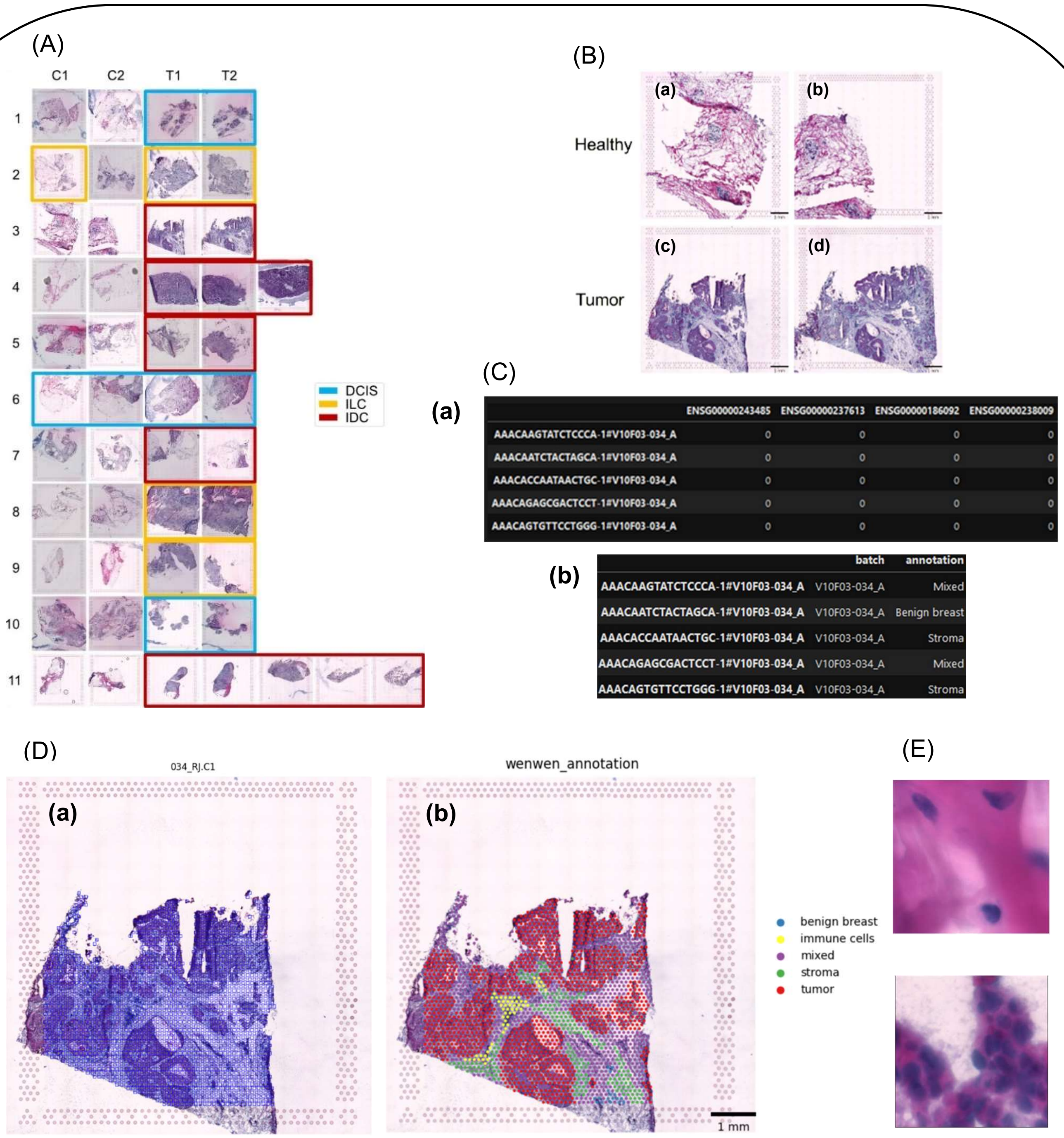
1. Data preprocessing and dataset

To study morphology and gene expression of tissues in benign and cancerous breast biopsy, a spatial transcriptomics dataset with data from eleven breast cancer patients with either ductal carcinoma in situ (DCIS), invasive lobular carcinoma (ILC) or invasive ductal carcinoma (IDC), was used (Figure 2A). For most analyses, four images from patient 3 were selected, two of which composed of healthy tissue and two sections containing tumor tissue (Figure 2B). A lot of analyses were also conducted only on the slide c of the patient 3.

For all the transcriptomics analysis, the transcriptomics data where pre-processed as explained in the materials and methods part.

The transcriptomics analysis was performed on different batch of different size. The data of each batch are presented in a dataset with X rows and 36601 columns, where X is the number of spots of the batch and 36601 is the number of genes. This allows to visualize the expression count of these 36601 genes for each spot (Figure 2C-a). Moreover, this dataframe is directly linked to an annotation dataframe where we can find the pathologists' annotations of each spot (Figure 2C-b). The barcodes allow to link these 2 dataframes.

For image analysis, image tiles of 65x65 µm were created from the spots located inside the tissue (Figure 2D, E). An expert pathologist has annotated each spot with a tissue type thanks to the morphological aspect of the images' tiles (Figure 2D-b).



	ENSG00000243485	ENSG00000237613	ENSG00000186092	ENSG00000238009
AAACAAGTATCTCCCA-1#V10F03-034_A	0	0	0	0
AAACAATCTACTAGCA-1#V10F03-034_A	0	0	0	0
AAACACCAATAACTGC-1#V10F03-034_A	0	0	0	0
AAACAGAGCGACTCCT-1#V10F03-034_A	0	0	0	0
AAACAGTGTTCCTGGG-1#V10F03-034_A	0	0	0	0

	batch	annotation
AAACAAGTATCTCCCA-1#V10F03-034_A	V10F03-034_A	Mixed
AAACAATCTACTAGCA-1#V10F03-034_A	V10F03-034_A	Benign breast
AAACACCAATAACTGC-1#V10F03-034_A	V10F03-034_A	Stroma
AAACAGAGCGACTCCT-1#V10F03-034_A	V10F03-034_A	Mixed
AAACAGTGTTCCTGGG-1#V10F03-034_A	V10F03-034_A	Stroma

Figure 2 A Complete ST breast cancer dataset. The dataset contains 48 images of 11 patients, who were diagnosed with ductal carcinoma in situ (DCIS), invasive lobular carcinoma (ILC) or invasive ductal carcinoma (IDC).
B Slides from patient 3 which were used for most analysis. Two images contained healthy tissue, and two images contained tumor tissue. The slide **c** was also used separately for some analyses. **C** Example of dataset head. **a** Gene expression matrix. **b** Annotation dataframe. **D** Example of a slide with ST spots. The spots are located 100 µm apart. **a** Images were tiled following the ST spot positions for image analyses. **b** Pathologist annotations on the slide. **E** Example of tiles from the slide a of the patient 3.

2. Identification of Tissue Types in Histopathology Slides Using Spatially Resolved Expression Data

Firstly, the transcriptomics analysis was performed on only one slide to get an overview of the results that we can get, while having easier analysis and less computational time.

A Random Forest (RF) was performed on the transcriptomics data to evaluate how accurate can we predict a tissue type with these data (see Figure 3a). A UMAP from these data was also plotted to observe how well can we segregate tissue type (see Figure 3b). The overall prediction accuracy of the model is 0.68, and we can observe that only the Tumor tissue type is well distinguished by the model (f1-score of 0.83, while other tissue types have a f1-score < 0.52). UMAP visualization reveals that Tumor is well segregated while other tissue types are mixed. This can explain why the model is not efficient. Moreover, the low quantity of IC, Stroma and BB tissue spots can explain the difficulty to classify them.

To see if a bigger dataset with more samples and spots would increase the accuracy of the model, the same approach was then performed on the 4 slides of the patient 3.

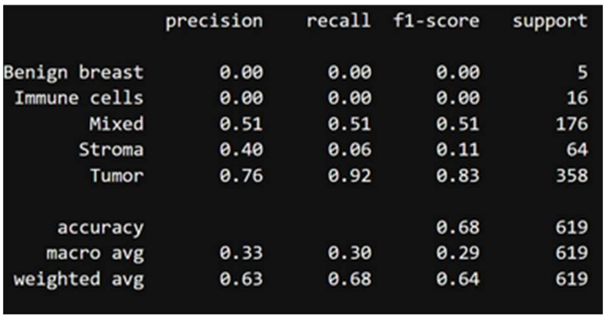
The results of the RF and the UMAP are shown in Annex 2. The overall prediction accuracy of the model is 0.72, which is better than for the slide 3C, and we can observe that both Tumor and Stroma tissue are well distinguished by the model (f1-score of 0.81 and 0.83, respectively, while other tissue types have a f1-score < 0.40). UMAP visualization reveals that Tumor and Stroma are well segregated while other tissue types are mixed. There is still a low quantity of IC, Fat and BB tissue spots which can explain the difficulty to classify them. Moreover, mixed spots are in between all other tissue types on the UMAP, certainly because of the heterogeneity of cells in mixed spots. This can also decrease the performance of the model.

To see if a dataset with more spots of Fat, IC, and BB would increase the accuracy of the model, the same approach was then performed on the 4 slides of the patient 3 and some slides of other patient containing a lot of spots from these tissues.

The results of the RF and the UMAP are shown in Annex 3. The overall prediction accuracy of the model is 0.49, which is worse than for the two previous models. However, we can observe that the tissue types identification is more balanced (f1-score between 0.39 and 0.69, except for IC which has still a f1-score of 0). UMAP visualization reveals spots from different tissue type are not segregated from each other. It is possibly due to a batch effect between all the slides from different patients that were used. Or maybe because of different types of tumor and different stages of cancer between patients. This could explain the low performance of this model.

To see if Mixed spots were decreasing the performance of models, the same approach was then performed on the 4 slides of the patient 3 without the Mixed spots.

The results of the RF and the UMAP are shown in Annex 4. The overall prediction accuracy of the model is 0.87, which is better than all the other transcriptomics models. We can observe that both Tumor and Stroma tissue are well distinguished by the model, but other tissue type are still not recognized (f1-score of 0.90 and 0.91, respectively, while other tissue types have a f1-score < 0.15). UMAP visualization reveals that Tumor and Stroma are well segregated while other tissues look more segregated than before. Despite a better segregation, there is still a low quantity of IC, Fat and BB tissue spots which can explain the difficulty to classify them.



```
               precision    recall  f1-score   support

Benign breast       0.00      0.00      0.00         5
 Immune cells       0.00      0.00      0.00        16
        Mixed       0.51      0.51      0.51       176
       Stroma       0.40      0.06      0.11        64
        Tumor       0.76      0.92      0.83       358

     accuracy                           0.68       619
    macro avg       0.33      0.30      0.29       619
 weighted avg       0.63      0.68      0.64       619
```

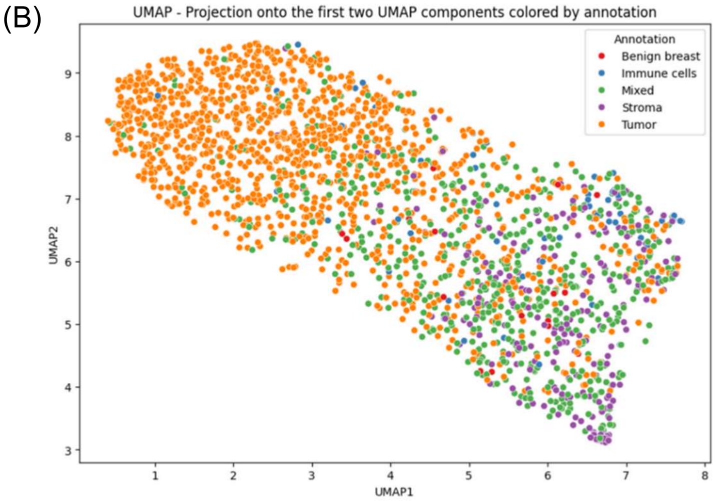


Figure 3 a Classification report of a RF performed on the slide 3C's transcriptomics data. The classification was performed between 2061 spots and 5 tissue type: Tumor, Mixed, IC, Stroma and BB (see in Annex 1). 1442 spots (70%) were used for training the model and 619 were used for testing the model (30%). **b** UMAP of the 2061 spots from the slide 3C (transcriptomics data).

3. Gene Expression Markers for the Characterization of Tissue Types

A DEA was conducted to identify genes that are significantly different between each tissue type, and that may be the ones that are used by the model to classify. The goal here was to see if the genes used by the model can be relevant biologically speaking.

The DEA was performed on patient 3 transcriptomics data (4 slides), with FDR adjusted P-value (Adjusted P-value < 0.05, Log2 fold-change > 1 or Log2 fold-change < -1). A total of 142 genes are up-regulated and a total of 182 genes are down-regulated in the dataset. The distribution of genes down or up regulated for each tissue type is available in Table 1. Most of the up regulation occurs in Tumor tissues while most of the down regulation occurs in Stroma tissues. Furthermore, the expression of tumor up-regulated genes is in strong opposition with the expression of the same genes in stromal tissues (Figure 4a). Similarly, the expression of stroma down-regulated genes is in strong opposition to their expression in tumor tissues (Figure 4b). These genes may be used by the model to distinguish Stroma from Tumor tissues. Moreover, Stroma tissues often have specific expression pattern in tumor microenvironment, due for example to fibroblast which activate in injurious and damaged tissue.[16]

A gene ontology enrichment analysis was further performed on the differential expressed genes (324 transcripts). The analysis highlighted enrichment of gene sets for each tissue type. For example, a lot of gene implied in immune function were found in up-regulated IC genes: B cell receptor signaling pathway, adaptive immune response (and more, see in Annex 5a). On the other hand, Tumor tissue appears to up-regulate genes implied in metabolic process and modeling of the immune response: fatty acid metabolic process, regulation of toll-like receptor signaling pathway (and more, see in Annex 5f). Increasing of fatty acid metabolic process is a common thing in cancer development and is often a sign of a bad prognostic.[17] Moreover, immune system process is also known to be enhanced in cancer.[17] For Stroma tissue, genes implied in cell activity and immune response were down regulated: mitochondrial ATP synthesis coupled proton transport, B cell receptor signaling pathway (and more, see in Annex 5g). The stromal cells are known to work as positive or negative regulators of tumor growth. Nevertheless, the primary function of fibroblasts is to respond to tissue injury and facilitate regenerative repair.[16] Therefore, an inhibition of Stroma tissue activity in a cancer context might increase tumor growth and decrease the prognostic. The full GO analysis is available in annex 5. In a general manner, genes that are differentially expressed and that may be used by the model for classification are relevant biologically speaking.

Tissue type	BB	Fat	IC	Mixed	Stroma	Tumor
Up regulated	7	3	12	7	1	112
Down regulated	4	3	0	0	166	9

Table 1 Distribution of genes down or up regulated for each tissue type

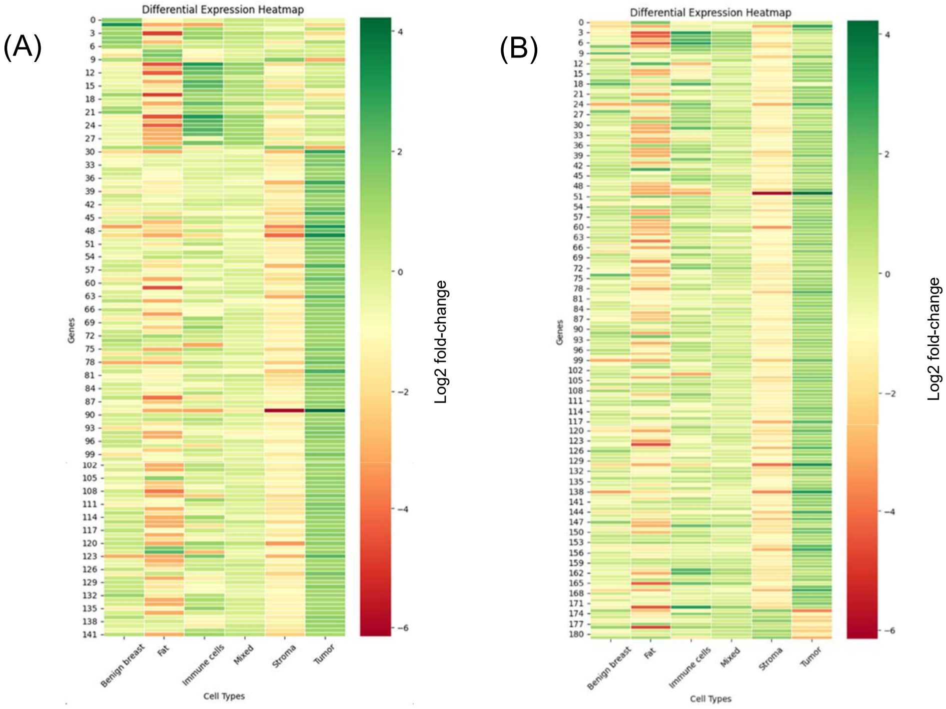
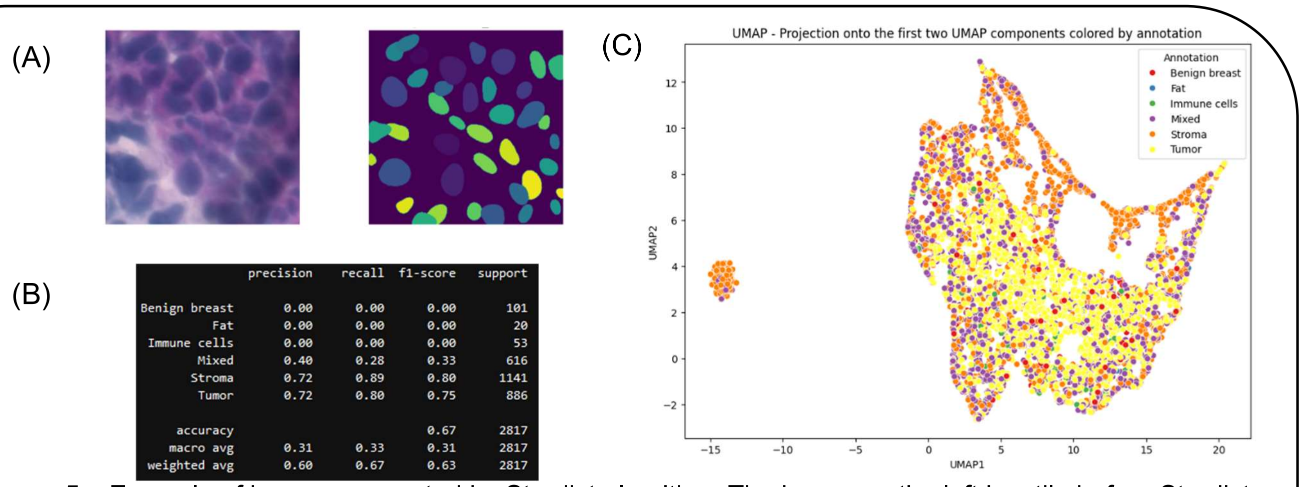


Figure 4 a Heatmap of the up-regulated genes in patient 3. All the genes that are up regulated in a least one tissue type are plotted on this graph. **B** Heatmap of the down-regulated genes in patient 3. All the genes that are down regulated in a least one tissue type are plotted on this graph.

4. Nucleus Features for Tumor Tissue Identification in Histopathological Images

Then, the goal was to evaluate whether histopathological image data would be efficient to identify different tissue types in breast cancer biopsy. The first method used was Stardist, which is an algorithm capable of identifying round object in an image, and so nuclei in histopathological images (Figure 5a). The model was trained and tested on image data of the 4 slides of the patient 3.

A RF analysis was conducted on the histopathological image data to evaluate the accuracy of predicting tissue types using these data (see Figure 5b). Additionally, a UMAP was generated to visualize how well the tissue types could be separated (see Figure 5c). The model's overall prediction accuracy is 0.67, with the Tumor and the Stroma tissues being the only ones clearly distinguished by the model (f1-score of 0.75 and 0.80, respectively, whereas other tissue types have f1-scores below 0.33). UMAP visualization shows that Tumor is clustered in the middle, while other tissue types are intermixed, which may explain the model's inefficiency. Since Stardist recognize nuclei, it may be easier to distinguish tumoral and non-tumoral tissue because their nuclei are quite different. However, for the non-tumoral tissue, it may be more difficult to spot nuclei differences. Furthermore, the low number of IC, Stroma, and BB tissue spots might contribute to the difficulty in classifying these types.



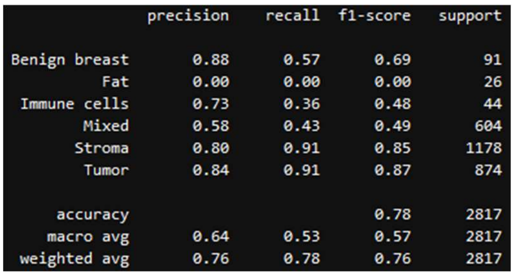
```
              precision    recall  f1-score   support

Benign breast      0.00      0.00      0.00       101
          Fat      0.00      0.00      0.00        20
 Immune cells      0.00      0.00      0.00        53
        Mixed      0.40      0.28      0.33       616
       Stroma      0.72      0.89      0.80      1141
        Tumor      0.72      0.80      0.75       886

     accuracy                          0.67      2817
    macro avg      0.31      0.33      0.31      2817
 weighted avg      0.60      0.67      0.63      2817
```

Figure 5 a Example of image segmented by Stardist algorithm**.** The image on the left is a tile before Stardist segmentation, the image on the right is after Stardist segmentation **b** Classification report of a RF performed on the image data (Stardist) of the 4 slides of the patient 3. The classification was performed between 9390 spots and 6 tissue type: Tumor, Mixed, IC, Stroma, BB and Fat (see in Annex 1). 6573 spots (70%) were used for training the model and 2817 were used for testing the model (30%). **c** UMAP of the 9390 spots from the 4 slides of the patient 4 (image data from KimiaNet).

5. Using Image Patterns to Identify Different Tissue Types in Histopathological Images

To evaluate whether histopathological image data would be efficient to identify different tissue types in breast cancer biopsy, the second method used was an FCNN pretrained on histopathological images called KimiaNet. The model was trained and tested on image data of the 4 slides of the patient 3.

A RF analysis was conducted on the histopathological image to evaluate the accuracy of predicting tissue types using these data (see Figure 6a). Additionally, a UMAP was generated to visualize how well the tissue types could be separated (see Figure 6b). The model's overall prediction accuracy is 0.78, which is the best overall prediction accuracy between all the previous models of the patient 3 (expect the transcriptomics model where Mixed spots where removed). The Tumor and Stroma tissue types are clearly distinguished by the model (f1-score of 0.87 and 0.85, respectively) whereas other tissue types have quite better f1-scores than for previous models (f1-score 0.69, 0.48, 0.49, for BB, IC and Mixed respectively). Only Fat tissue is not recognized in this model (f1-score = 0). UMAP visualization shows that tissues are well segregated, but it may have a batch effect between slides containing mostly Stroma and slides containing mostly Tumor. Moreover, the model distinguishes two different types of Tumor tissue, which may be also a batch effect or just different tumor stages. Furthermore, the low number of Fat spots and the difficulty to stain hydrophobic tissue might contribute to the difficulty in classifying them.



```
              precision    recall  f1-score   support

Benign breast       0.88      0.57      0.69        91
          Fat       0.00      0.00      0.00        26
 Immune cells       0.73      0.36      0.48        44
        Mixed       0.58      0.43      0.49       604
       Stroma       0.80      0.91      0.85      1178
        Tumor       0.84      0.91      0.87       874

     accuracy                           0.78      2817
    macro avg       0.64      0.53      0.57      2817
 weighted avg       0.76      0.78      0.76      2817
```

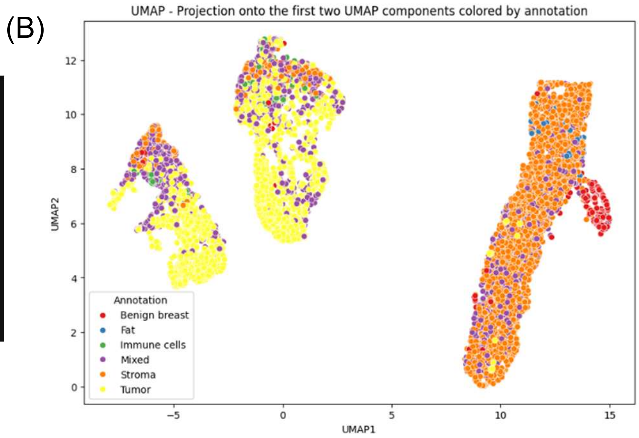


Figure 6 a Classification report of a RF performed on the image data (KimiaNet) of the 4 slides of the patient 3. The classification was performed between 9390 spots and 6 tissue type: Tumor, Mixed, IC, Stroma, BB and Fat (see in Annex 1). 6573 spots (70%) were used for training the model and 2817 were used for testing the model (30%). **b** UMAP of the 9390 spots from the 4 slides of the patient 4 (image data from KimiaNet).

6. Enhancing Tissue Type Identification Through Integration of Spatially Resolved Expression Data and Histopathological Features

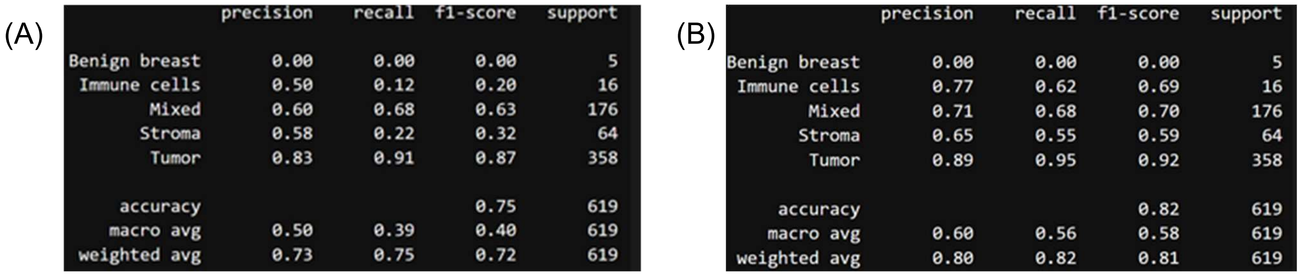
In this part, the goal was to combine image and gene data with multiple intelligence processing algorithms (Multimodal AI) to achieve higher performances. All the models in this section were conducted on data of the slide C from the patient 3.

Firstly, a simple concatenation of Stardist Image Data (SID) and Transcriptomics Data (TD) didn't improve the overall prediction accuracy of the model compared to TD alone (Table 2, Concatenation Stardist). Similarly, the concatenation of KimiaNet Image Data (KID) and TD didn't improved the overall prediction accuracy of the model compared to KID alone (Table 2, Concatenation KimiaNet, and Figure 7a). Using a CNN to combine SID and TD improved the overall prediction accuracy of 3% compared to TD alone (Table 2, CNN Stardist). Using a stacking method to combine SID and TD improved the overall prediction accuracy of 5% compared to TD alone (Table 2, Stacking Stardist). The stacking was composed of RF, gradient boosting, K-nearest neighbor and support vector machine as base estimator and had a logistigtic regression as a meta-estimator. The ConGI tool provided an overall prediction accuracy of 0.75 (Table 2, ConGI) and the identification of different tissue types was balanced, with 0.2 for IC and 0.32 for Stroma, despite the low amount of spots of these tissues (Figure 7b).

In a general manner, combining gene and image data improve the classification of tissue types. However, it is challenging to improve drastically the accuracy with these models.

Model	TD	SID	KID	Concatenation Stardist	Concatenation KimiaNet	CNN Stardist	Stacking Stardist	ConGI
Accuracy	0.68	0.63	0.82	0.68	0.82	0.71	0.73	0.75

Table 2 Summary of the overall prediction accuracy of different models trained on slide C from the patient 3



(A)

```
               precision    recall  f1-score   support

Benign breast       0.00      0.00      0.00         5
 Immune cells       0.50      0.12      0.20        16
        Mixed       0.60      0.68      0.63       176
       Stroma       0.58      0.22      0.32        64
        Tumor       0.83      0.91      0.87       358

     accuracy                           0.75       619
    macro avg       0.50      0.39      0.40       619
 weighted avg       0.73      0.75      0.72       619
```

(B)

```
               precision    recall  f1-score   support

Benign breast       0.00      0.00      0.00         5
 Immune cells       0.77      0.62      0.69        16
        Mixed       0.71      0.68      0.70       176
       Stroma       0.65      0.55      0.59        64
        Tumor       0.89      0.95      0.92       358

     accuracy                           0.82       619
    macro avg       0.60      0.56      0.58       619
 weighted avg       0.80      0.82      0.81       619
```

Figure 7 Multimodal AI trained on slide C from the patient 3 **a** Model using TD and KID. The accuracy is the same as KID only, but using KID in a model still give the best performance. **b** Model using ConGI. The accuracy of this model is better than every model except models using KID.

7. Recognize Breast Cancer Tissue Types with Unsupervised Machine Learning Using Spatial Transcriptomics Data and Histopathological Images

TD and Histopathological Images can also be used for clustering tasks in a tissue. The goal in this section is to evaluate how accurately different models can achieve clustering.

All the clustering tasks were conducted on data of the slide C from the patient 3. TD, SID, KID and ConGI performances were tested and compared. ConGI, TD, and KID have approximately the same clustering performance: ARI of 0.15, 0.16, and 0.14, respectively (Figure 8 a, c and e). On the other hand, SID is not reliable for clustering tasks with an ARI of 0.04 (Figure 8d). ConGI and KID seem to recognize two different types of Tumor tissue (in orange and in green for ConGI, and in yellow and purple for KID), the Stroma tissue (in red and in blue, respectively), and a part of the Mixed tissue (in blue and in green, respectively) (Figure 8 a, b and e). Other tissue are not represented by the clustering of ConGI and KID.

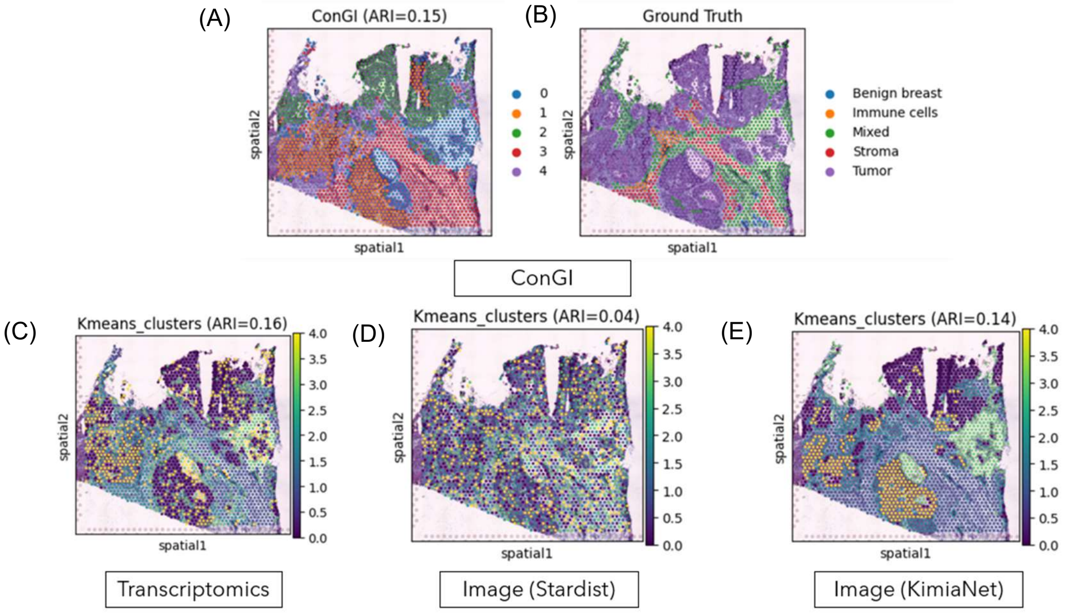


Figure 8 Clustering by different data types and models. Clustering was conducted on slide C from the patient 3. Performance of the clustering was evaluated with the ARI metrics. Kmeans clusters was used for TD, SID and KID. Mclust was used for ConGI's model. **a** ConGI model **b** Pathologist's annotations **c** Transcriptomics data **d** Stardist image data **e** KimiaNet image data.

TD seem to recognize also two types of tumor tissue (in purple and in yellow) and to separate quite well Mixed and Stroma tissue (green and blue, respectively). However, TD model is not able to recognize IC and BB tissue (Figure 8 b and c).

UMAP representations of the clustering were also plotted for the different data types and models (Annex 6). On these UMAP we can observe that for TD, ConGI and KID the data are quite well segregated for each tissue types. However, there is always data intermixed in each model, which may explain the difficulty to cluster the different tissue types. We can also observe that there are for all these models two clusters corresponding to the tumor annotated tissues. This may be caused by tumor heterogeneity in the sample. UMAP of SID model revealed why the clustering has a low performance since all the data are intermixed.

IV. Discussion

In this study, several models have been tested on 4 tissue section slides of the same patient to evaluate the capacity of ML and each data type to be used for tissue type identification. Firstly, supervised ML was performed separately on transcriptomics and histopathological image data. Classification results gave the best performance for KID, then TD and finally SID (overall accuracy of 0.78, 0.72, 0.68, respectively. Figure 5a, 6a, Annex 2a). In the same way, the f1-score for cancer tissue classification gave the best performance for KID, then TD and finally SID (0.87, 0.81, 0.75, respectively. Figure 5a, 6a, Annex 2a). These results overall demonstrated the power of ST data to classify tissue types in breast cancer tissue sections, and more specifically to classify cancer regions. Distinguishing the different non tumor tissue types remain challenging, possibly because of the low number of non-tumoral spots in the slides used in this study or because of the heterogeneity of the tumor microenvironment (regardless of the method, none of the tissue type, except Stroma, got an f1-score higher than 0.80). Fat tissue spots also remain a challenge, possibly because of their hydrophobic characteristics which might make the RNA extraction difficult during the ST (f1-score of 0.00 for all the method except for the TD model enriched in non-tumoral spots. See Annex 3.a).

Then, integration of TD and histopathological image data was tested in several models for classification tasks, on data of one slide only. The best model was the one using KID. However, adding the TD didn't improve the overall accuracy compared to the KID model without TD (Table 2). On the other hand, several models improved the overall accuracy compared to TD or SID alone. Indeed, CNN, Stacking and ConGI models gave a better overall accuracy with 0.71, 0.73 and 0.75, respectively (Table 2). ConGI architecture is very promising because it improved the most the overall accuracy compared to the single modal TD model. However, ConGI is using a DenseNet CNN, which is train on non-biological images and is less performant than KimiaNet for histopathological images. It would be interesting to try to reproduce the ConGI architecture but with KimiaNet as a CNN instead of DenseNet. In a general manner, KimiaNet remains the best model for supervised ML and classification tasks.

Further analysis employed unsupervised classification of ST spots for identifying tissue types. The resulting five distinct classes didn't fit to the five expert annotated regions. Indeed, KID, TD and ConGI models were able to identify tumor tissue in two different clusters, and to identify Stroma and Mixed tissue, but BB and IC tissue were not identified (Figure 8). The clustering performance of this study's unsupervised models (average ARI of 0.15) was less accurate compared to the original ConGI study conducted on the HER+ dataset (average ARI of 0.36)[18]. The low number of BB and IC ST spots used in this study might make it difficult to cluster them. In addition, the HER+ dataset has expert annotation on the tumor types, whereas our data have expert annotation on tissue types which are intermixed, and which might add variances in the data and complicate the clustering task.

The DEA and the GO analysis showed that significant differentially expressed genes for each tissue types could be relevant biologically (Figure 4 and Annex 5). This may be the same gene expression signatures that the transcriptomics models are using for classification. This suggest that the results might be reproducible between different breast cancer samples. However, this hypothesis needs to be validated by repeating the analysis on other samples and especially sample with different tumor types or stages that might have different gene expression signatures.

The transcriptomics model enriched in non-tumoral tissues showed that combining different samples could allow to identify more accurately the non-tumoral tissue (Annex 3). However, it also showed that it decreases the tumor and stroma detection. This could be due to a batch effect or some tumor heterogeneity between different samples. A DEA and GO analysis should be done on this model to evaluate whether the difference between sample have a biological explanation or if it is a technical issue.

The transcriptomics model without the Mixed annotated spots showed better results than any other transcriptomics model (accuracy of 0.87). Mixed spots appeared to have an expression pattern at the border between all the other tissue types. This could be easily explained because these Mixed annotated spots are basically a mixture of several tissue types in a spot. To improve transcriptomics model accuracy, a challenge would be to identify better these Mixed spots. One could use single cell data from breast cancer tissue to complement the ST data. Indeed, single cell data could be used to analyze the expression pattern of these Mixed spots with deconvolution methods[19], which will give the percentage of each cell type in each ST spot. We could then reassign a new annotation to these spots following the major cell type of each Mixed spot. Moreover, expression data from ST have a shallow sequencing, while single cell data allow to get a deeper sequencing.

Finally, using TD to supplement histopathological images for identification of tissue in breast cancer ST slides is promising. However, the performance of multimodal AI models does not outperform single modal AI image model for now. Therefore, there is a need to develop more efficient models capable of combine image data and TD. For developing a tool capable of diagnosing breast cancer, the ST data do not seem to be necessary for now. All the more so since the ST data production is more expensive than just a staining to obtain histopathological images.

V. Conclusion

This study reported the use of ML on a ST dataset to identify different tissue in breast cancer sections. Firstly, the performance of single modal AI of transcriptomics data and image data were evaluated separately for supervised and unsupervised ML. Then, the performance of multimodal AI, which is combining image and transcriptomics data was evaluated for supervised and unsupervised ML. Finally, the gene expressions signatures were analyzed to evaluate the biological relevance of the results. Computer-guided detection of tumor and non-tumoral tissues in breast cancer sections using ST data is a great hope to assist experts pathologists in diagnoses of cancer.

VI. Bibliography

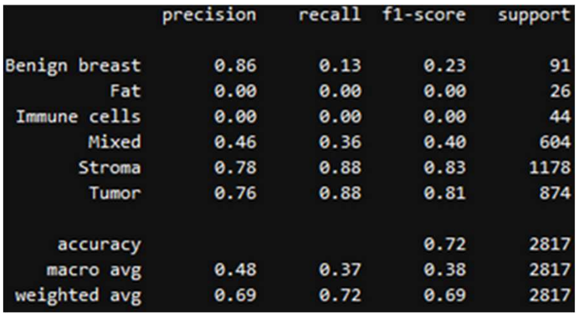
1. Breast cancer. https://www.who.int/news-room/fact-sheets/detail/breast-cancer.
2. Cancers du sein : les prévenir et les détecter tôt - Dépistage du cancer du sein. https://www.e-cancer.fr/Comprendre-prevenir-depister/Se-faire-depister/Depistage-du-cancer-du-sein/Prevenir-et-depister-tot.
3. Breast cancer - Diagnosis and treatment - Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/breast-cancer/diagnosis-treatment/drc-20352475.
4. Breast Cancer Diagnosis - National Breast Cancer Foundation. https://www.nationalbreastcancer.org/breast-cancer-diagnosis/.
5. Breast Cancer Diagnosis, Tests and Early Detection. *City of Hope* https://www.cancercenter.com/cancer-types/breast-cancer/diagnosis-and-detection (2018).
6. Fumagalli, C. & Barberis, M. Breast Cancer Heterogeneity. *Diagnostics (Basel)* **11**, 1555 (2021).
7. Yoosuf, N., Navarro, J. F., Salmén, F., Ståhl, P. L. & Daub, C. O. Identification and transfer of spatial transcriptomics signatures for cancer diagnosis. *Breast Cancer Research* **22**, 6 (2020).
8. Machine learning, explained | MIT Sloan. https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained (2024).
9. Greener, J. G., Kandathil, S. M., Moffat, L. & Jones, D. T. A guide to machine learning for biologists. *Nat Rev Mol Cell Biol* **23**, 40–55 (2022).
10. Acosta, J. N., Falcone, G. J., Rajpurkar, P. & Topol, E. J. Multimodal biomedical AI. *Nat Med* **28**, 1773–1784 (2022).
11. Scanpy – Single-Cell Analysis in Python. *Scanpy* https://scanpy.readthedocs.io/en/stable/index.html.
12. DAVID Functional Annotation Bioinformatics Microarray Analysis. https://david.ncifcrf.gov/.
13. Identifying Cells And Epithelia Lab. https://medcell.org/tbl/identifying_cells_and_epithelia/reading.php.
14. stardist/stardist. StarDist (2024).
15. KimiaLabMayo/KimiaNet. Kimia Lab (2024).
16. Kalluri, R. The biology and function of fibroblasts in cancer. *Nat Rev Cancer* **16**, 582–598 (2016).
17. Creighton, C. J. Gene expression profiles in cancers and their therapeutic implications. *Cancer J* **29**, 9–14 (2023).
18. Zeng, Y. *et al.* Identifying spatial domain by adapting transcriptomics with histology through contrastive learning. *Briefings in Bioinformatics* **24**, bbad048 (2023).
19. Zeng, Z., Li, Y., Li, Y. & Luo, Y. Statistical and machine learning methods for spatially resolved transcriptomics data analysis. *Genome Biology* **23**, 83 (2022).

Annex:

Tissue type		Tumor	Mixed	IC	Stroma	BB	Fat
Spot number	Slide 3C	1209	565	57	213	17	0
	Patient 3	2917	2046	162	3873	322	70
	Patient 3 + specifics annotations	5183	7334	329	6269	4107	3556
	Patient 3 without mixed annotations	2917	0	162	3873	322	70

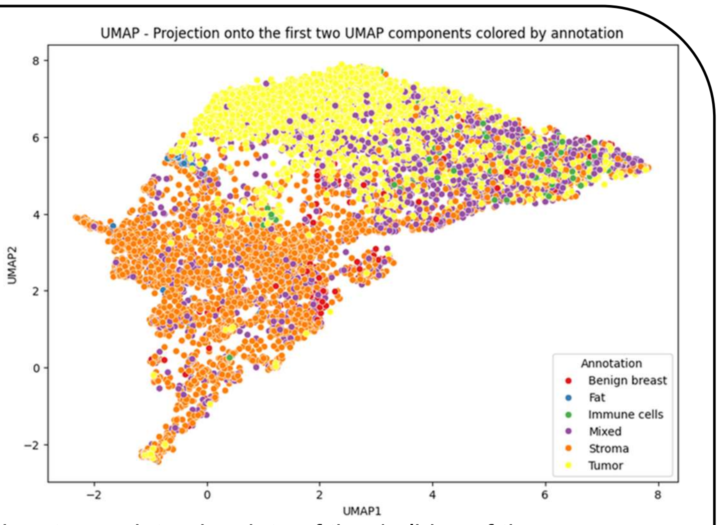
Annex 1. Annotation distribution in different samples

(A)



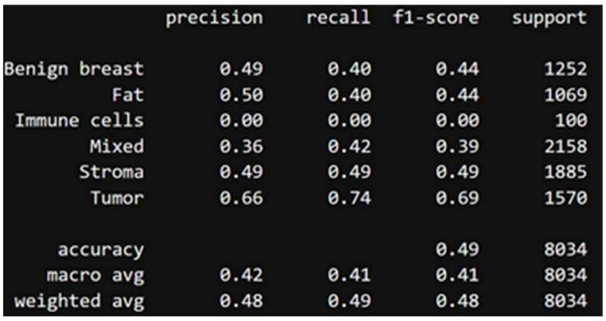
```
               precision    recall  f1-score   support

Benign breast       0.86      0.13      0.23        91
          Fat       0.00      0.00      0.00        26
 Immune cells       0.00      0.00      0.00        44
        Mixed       0.46      0.36      0.40       604
       Stroma       0.78      0.88      0.83      1178
        Tumor       0.76      0.88      0.81       874

     accuracy                           0.72      2817
    macro avg       0.48      0.37      0.38      2817
 weighted avg       0.69      0.72      0.69      2817
```

(B)



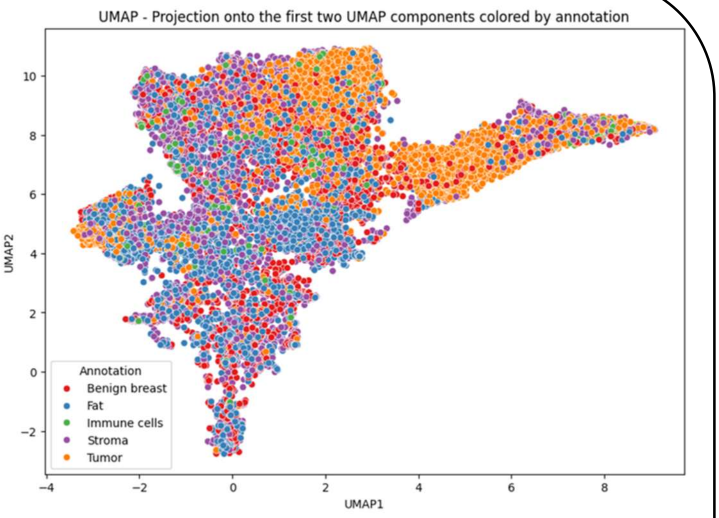
Annex 2. a Classification report of a RF performed on transcriptomics data of the 4 slides of the patient 3. The classification for the patient 3 was performed between 9390 spots and 6 tissue type: Tumor, Mixed, IC, Stroma, BB and Fat (see in Annex 1). 6573 spots (70%) were used for training the model and 2817 were used for testing the model (30%). **b** UMAP of the 9390 spots from the 4 slides of the patient 4.

(A)



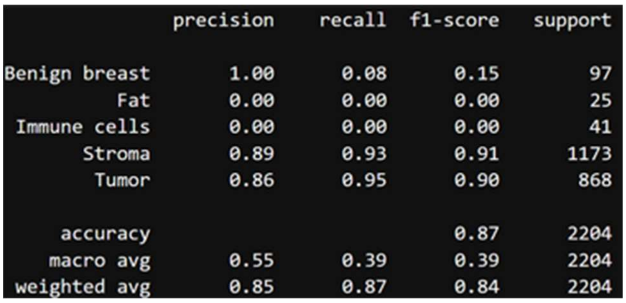
```
               precision    recall  f1-score   support

Benign breast       0.49      0.40      0.44      1252
          Fat       0.50      0.40      0.44      1069
 Immune cells       0.00      0.00      0.00       100
        Mixed       0.36      0.42      0.39      2158
       Stroma       0.49      0.49      0.49      1885
        Tumor       0.66      0.74      0.69      1570

     accuracy                           0.49      8034
    macro avg       0.42      0.41      0.41      8034
 weighted avg       0.48      0.49      0.48      8034
```

(B)



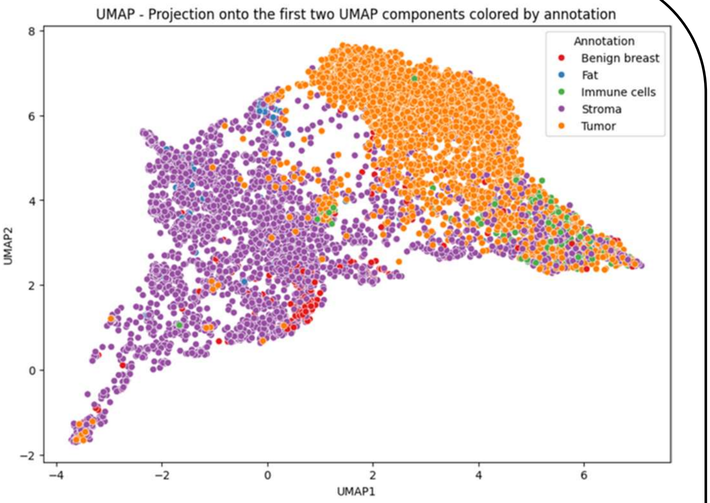
Annex 3. a Classification report of a RF performed on transcriptomics data of the 4 slides of the patient 3 and some slides of other patients containing a lot of Fat, BB and IC spots. The classification for this model was performed between 26 778 spots and 6 tissue type: Tumor, Mixed, IC, Stroma, BB and Fat (see in Annex 1). 18 774 spots (70%) were used for training the model and 8034 were used for testing the model (30%). **b** UMAP of the 26 778 spots from the same model.

(A)



```
               precision    recall  f1-score   support

Benign breast       1.00      0.08      0.15        97
          Fat       0.00      0.00      0.00        25
 Immune cells       0.00      0.00      0.00        41
       Stroma       0.89      0.93      0.91      1173
        Tumor       0.86      0.95      0.90       868

     accuracy                           0.87      2204
    macro avg       0.55      0.39      0.39      2204
 weighted avg       0.85      0.87      0.84      2204
```

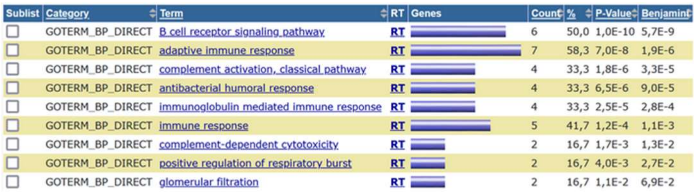
(B)



Annex 4. a Classification report of a RF performed on transcriptomics data of the 4 slides of the patient 3 and without Mixed spots. The classification for this model was performed between 7344 spots and 5 tissue type: Tumor, IC, Stroma, BB and Fat (see in Annex 1). 5140 spots (70%) were used for training the model and 2204 were used for testing the model (30%). **b** UMAP of the 7344 spots from the same model.

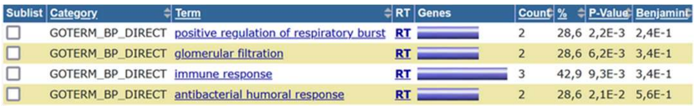
Annex 5: Gene Ontology analysis with DAVID

For up regulated genes:



Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
□	GOTERM_BP_DIRECT	B cell receptor signaling pathway	RT		6	50,0	1,0E-10	5,7E-9
□	GOTERM_BP_DIRECT	adaptive immune response	RT		7	58,3	7,0E-8	1,9E-6
□	GOTERM_BP_DIRECT	complement activation, classical pathway	RT		4	33,3	1,8E-6	3,3E-5
□	GOTERM_BP_DIRECT	antibacterial humoral response	RT		4	33,3	6,5E-6	9,0E-5
□	GOTERM_BP_DIRECT	immunoglobulin mediated immune response	RT		4	33,3	2,5E-5	2,8E-4
□	GOTERM_BP_DIRECT	immune response	RT		5	41,7	1,2E-4	1,1E-3
□	GOTERM_BP_DIRECT	complement-dependent cytotoxicity	RT		2	16,7	1,7E-3	1,3E-2
□	GOTERM_BP_DIRECT	positive regulation of respiratory burst	RT		2	16,7	4,0E-3	2,7E-2
□	GOTERM_BP_DIRECT	glomerular filtration	RT		2	16,7	1,1E-2	6,9E-2

a Immune Cells. A lot of immune genes.

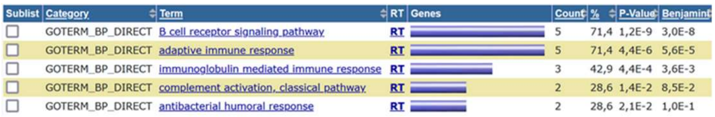


Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
□	GOTERM_BP_DIRECT	positive regulation of respiratory burst	RT		2	28,6	2,2E-3	2,4E-1
□	GOTERM_BP_DIRECT	glomerular filtration	RT		2	28,6	6,2E-3	3,4E-1
□	GOTERM_BP_DIRECT	immune response	RT		3	42,9	9,3E-3	3,4E-1
□	GOTERM_BP_DIRECT	antibacterial humoral response	RT		2	28,6	2,1E-2	5,6E-1

b Benign Breast. Biological Process not really linked to the benign breast cell type. BB are quite active in immune response to get a healthy milk, so the two last genes make sense. There is also in the breast in general a lot of immune cells next to benign breast cells. Also, the first gene is a gene common for cell activity in general. However, the is no lactation genes (CSN2 = casein, LALBA = protein essential for synthesis of lactose), or hormonal (ESR1 = estrogen receptor, PGR = progesterone receptor), or growth genes (ERBB2, EGFR), or differentiation genes (GATA3, FOXA1).

c Fat. No cluster for significant genes. But genes for apolipoprotein D (multi-ligand, multi-function protein that is involved lipid trafficking, food intake, inflammation, antioxidative response and development and in different types of cancers) and secretogoblin (regulation function, homeostasis, repair, inflammation, disease).
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8011330/

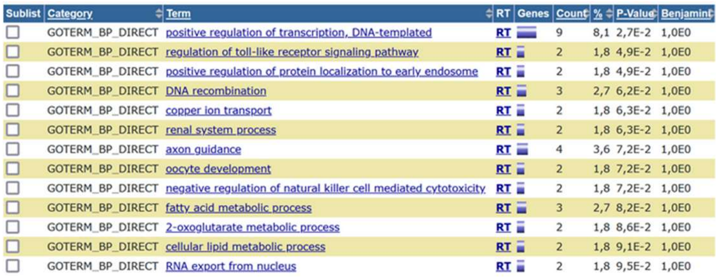
https://www.sciencedirect.com/topics/neuroscience/secretoglobin



Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
□	GOTERM_BP_DIRECT	B cell receptor signaling pathway	RT		5	71,4	1,2E-9	3,0E-8
□	GOTERM_BP_DIRECT	adaptive immune response	RT		5	71,4	4,4E-6	5,6E-5
□	GOTERM_BP_DIRECT	immunoglobulin mediated immune response	RT		3	42,9	4,4E-4	3,6E-3
□	GOTERM_BP_DIRECT	complement activation, classical pathway	RT		2	28,6	1,4E-2	8,5E-2
□	GOTERM_BP_DIRECT	antibacterial humoral response	RT		2	28,6	2,1E-2	1,0E-1

d Mixed. A lot of immune genes. Maybe a lot of immune cells are in mixed parts of the tissue. Or maybe that cells in general in cancer tissue are expressing more immune genes.

e Stroma. Gene for apolipoprotein D.

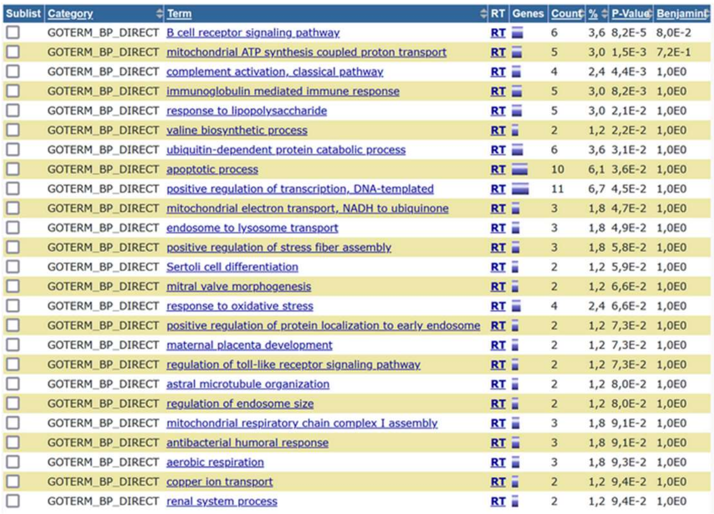


Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
□	GOTERM_BP_DIRECT	positive regulation of transcription, DNA-templated	RT		9	8,1	2,7E-2	1,0E0
□	GOTERM_BP_DIRECT	regulation of toll-like receptor signaling pathway	RT		2	1,8	4,9E-2	1,0E0
□	GOTERM_BP_DIRECT	positive regulation of protein localization to early endosome	RT		2	1,8	4,9E-2	1,0E0
□	GOTERM_BP_DIRECT	DNA recombination	RT		3	2,7	6,2E-2	1,0E0
□	GOTERM_BP_DIRECT	copper ion transport	RT		2	1,8	6,3E-2	1,0E0
□	GOTERM_BP_DIRECT	renal system process	RT		2	1,8	6,3E-2	1,0E0
□	GOTERM_BP_DIRECT	axon guidance	RT		4	3,6	7,2E-2	1,0E0
□	GOTERM_BP_DIRECT	oocyte development	RT		2	1,8	7,2E-2	1,0E0
□	GOTERM_BP_DIRECT	negative regulation of natural killer cell mediated cytotoxicity	RT		2	1,8	7,2E-2	1,0E0
□	GOTERM_BP_DIRECT	fatty acid metabolic process	RT		3	2,7	8,2E-2	1,0E0
□	GOTERM_BP_DIRECT	2-oxoglutarate metabolic process	RT		2	1,8	8,6E-2	1,0E0
□	GOTERM_BP_DIRECT	cellular lipid metabolic process	RT		2	1,8	9,1E-2	1,0E0
□	GOTERM_BP_DIRECT	RNA export from nucleus	RT		2	1,8	9,5E-2	1,0E0

f Tumor. Some immune genes. Genes that increase transcription. Metabolic process gene.

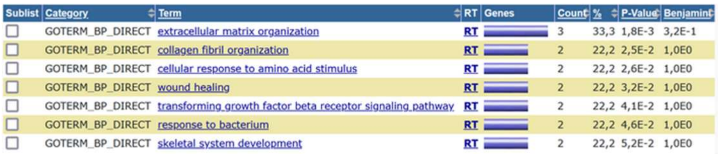
Increase transcription can be activation of transcription factor that increase the cell growth (e.g. transcription of growth factor). It also depends on the stage of cancer.

For down regulated genes:



Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamin
☐	GOTERM_BP_DIRECT	B cell receptor signaling pathway	RT		6	3,6	8,2E-5	8,0E-2
☐	GOTERM_BP_DIRECT	mitochondrial ATP synthesis coupled proton transport	RT		5	3,0	1,5E-3	7,2E-1
☐	GOTERM_BP_DIRECT	complement activation, classical pathway	RT		4	2,4	4,4E-3	1,0E0
☐	GOTERM_BP_DIRECT	immunoglobulin mediated immune response	RT		5	3,0	8,2E-3	1,0E0
☐	GOTERM_BP_DIRECT	response to lipopolysaccharide	RT		5	3,0	2,1E-2	1,0E0
☐	GOTERM_BP_DIRECT	valine biosynthetic process	RT		2	1,2	2,2E-2	1,0E0
☐	GOTERM_BP_DIRECT	ubiquitin-dependent protein catabolic process	RT		6	3,6	3,1E-2	1,0E0
☐	GOTERM_BP_DIRECT	apoptotic process	RT		10	6,1	3,6E-2	1,0E0
☐	GOTERM_BP_DIRECT	positive regulation of transcription, DNA-templated	RT		11	6,7	4,5E-2	1,0E0
☐	GOTERM_BP_DIRECT	mitochondrial electron transport, NADH to ubiquinone	RT		3	1,8	4,7E-2	1,0E0
☐	GOTERM_BP_DIRECT	endosome to lysosome transport	RT		3	1,8	4,9E-2	1,0E0
☐	GOTERM_BP_DIRECT	positive regulation of stress fiber assembly	RT		3	1,8	5,8E-2	1,0E0
☐	GOTERM_BP_DIRECT	Sertoli cell differentiation	RT		2	1,2	5,9E-2	1,0E0
☐	GOTERM_BP_DIRECT	mitral valve morphogenesis	RT		2	1,2	6,6E-2	1,0E0
☐	GOTERM_BP_DIRECT	response to oxidative stress	RT		4	2,4	6,6E-2	1,0E0
☐	GOTERM_BP_DIRECT	positive regulation of protein localization to early endosome	RT		2	1,2	7,3E-2	1,0E0
☐	GOTERM_BP_DIRECT	maternal placenta development	RT		2	1,2	7,3E-2	1,0E0
☐	GOTERM_BP_DIRECT	regulation of toll-like receptor signaling pathway	RT		2	1,2	7,3E-2	1,0E0
☐	GOTERM_BP_DIRECT	astral microtubule organization	RT		2	1,2	8,0E-2	1,0E0
☐	GOTERM_BP_DIRECT	regulation of endosome size	RT		2	1,2	8,0E-2	1,0E0
☐	GOTERM_BP_DIRECT	mitochondrial respiratory chain complex I assembly	RT		3	1,8	9,1E-2	1,0E0
☐	GOTERM_BP_DIRECT	antibacterial humoral response	RT		3	1,8	9,1E-2	1,0E0
☐	GOTERM_BP_DIRECT	aerobic respiration	RT		3	1,8	9,3E-2	1,0E0
☐	GOTERM_BP_DIRECT	copper ion transport	RT		2	1,2	9,4E-2	1,0E0
☐	GOTERM_BP_DIRECT	renal system process	RT		2	1,2	9,4E-2	1,0E0

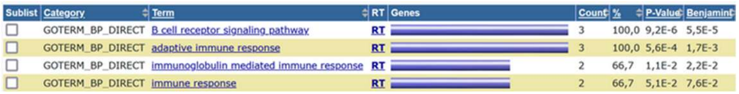
g Stroma. A lot of immune response genes are inhibited. Mitochondrial and cell activity genes are also down regulated. It seems that stromal cells activity and action in immune response are inhibited.



Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamin
☐	GOTERM_BP_DIRECT	extracellular matrix organization	RT		3	33,3	1,8E-3	3,2E-1
☐	GOTERM_BP_DIRECT	collagen fibril organization	RT		2	22,2	2,5E-2	1,0E0
☐	GOTERM_BP_DIRECT	cellular response to amino acid stimulus	RT		2	22,2	2,6E-2	1,0E0
☐	GOTERM_BP_DIRECT	wound healing	RT		2	22,2	3,2E-2	1,0E0
☐	GOTERM_BP_DIRECT	transforming growth factor beta receptor signaling pathway	RT		2	22,2	4,1E-2	1,0E0
☐	GOTERM_BP_DIRECT	response to bacterium	RT		2	22,2	4,6E-2	1,0E0
☐	GOTERM_BP_DIRECT	skeletal system development	RT		2	22,2	5,2E-2	1,0E0

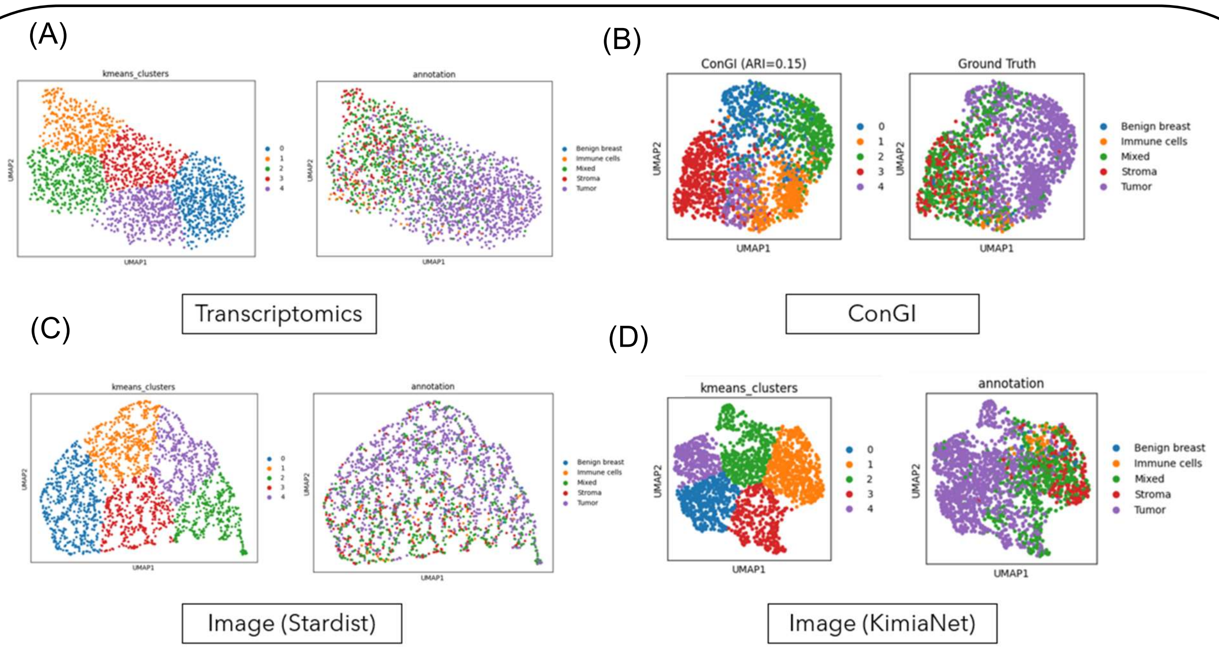
h Tumor. Organization genes are inhibited.

i Benign breast. No Biological process cluster. Genes for keratin (common epithelial marker), immunoglobulin, glucosidase, secretoglobin.



Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamin
☐	GOTERM_BP_DIRECT	B cell receptor signaling pathway	RT		3	100,0	9,2E-6	5,5E-5
☐	GOTERM_BP_DIRECT	adaptive immune response	RT		3	100,0	5,6E-4	1,7E-3
☐	GOTERM_BP_DIRECT	immunoglobulin mediated immune response	RT		2	66,7	1,1E-2	2,2E-2
☐	GOTERM_BP_DIRECT	immune response	RT		2	66,7	5,1E-2	7,6E-2

j Fat. Immune genes inhibited.



Annex 6 UMAP clustering representations for different data types and models. Clustering was conducted on slide C from the patient 3. Kmeans clusters was used for TD, SID and KID. Mclust was used for ConGI's model. For each model, annotation from the clustering algorithm are shown on the left UMAP, while pathologist's annotations are shown on the right UMAP. **a** Transcriptomics data **b** ConGI **c** Stardist image data **d** KimiaNet image data.

Bilan de compétence :

Qu'avez-vous tiré de votre stage ?	J'ai grandement amélioré mes compétences en communications scientifiques et surtout en anglais au travers de plusieurs présentations sur mes résultats devant l'équipe. J'ai appris à gérer une machine virtuelle et des données contenues sur un serveur. J'ai découvert le domaine du Machine Learning et j'ai pu implémenter plusieurs codes permettant d'identifier les tissus dans des sections mammaires cancéreuses. Je me suis documenté sur le cancer du sein et son fonctionnement biologique et sur les micro-environnements de tumeurs. J'ai découvert plusieurs techniques d'omique permettant d'étudier le cancer comme la transcriptomique spatiale et le single cell RNA sequencing. J'ai appris à gérer et à pré-traiter des données de transcriptomiques spatiales et à faire de l'analyse de ces données comme avec des analyses de DEA ou de GO. J'ai appris à gérer un espace GitLab afin de rendre mon code accessible à tous.
Qu'avez-vous apporté à la structure d'accueil ?	J'ai exploré le dataset de transcriptomique spatiale et j'ai essayé plusieurs méthodes novatrices de machine learning comme l'IA multimodale et le FCNN KimiaNet. J'ai répertorié tout mon travail sur le GitLab du laboratoire afin que ce que j'ai produit puisse être réutiliser ou étudier en cas de besoin.
Qu'avez-vous acquis ?	Des compétences sociales, computationnelles et biologiques.
Avez-vous atteint les objectifs que vous vous étiez fixés ?	Mon plan d'étude a été mené à bout. Cependant j'aurais aimé poursuivre l'étude sur l'IA multimodale et notamment essayer d'implémenter KimiaNet en tant que CNN permettant l'analyse des images dans ConGI (car les performances obtenues avec ConGI étaient moins bien qu'espérées).
Comment avez-vous géré votre temps ? Estimez-vous que vous avez réussi ?	En début de stage, j'ai commencé par me renseigner sur la transcriptomique spatiale et sur le machine learning en lisant des articles scientifiques. J'ai aussi effectué des tutoriels sur les modules python permettant de gérer les données de ST et d'utiliser du ML. J'ai effectué un plan de travail de ce que je voulais accomplir pendant le stage. Puis j'ai commencé par faire de l'analyse des images histopathologiques. J'ai ensuite enchainé par les données de ST et j'ai fini par de l'IA multimodale. J'ai ensuite utilisé les 3 dernières semaines de stage pour écrire mon rapport et organiser ma soutenance

	de stage. J’estime que j’ai réussi à faire la plupart des choses que je voulais.
Quels outils maîtrisez-vous davantage maintenant qu’au début de votre stage ?	Les serveurs, la machine virtuelle, le terminal bash, gitlab, python et les modules pour les données de ST (anndata, scanpy) et de ML (sklearn, tensorflow).
Souhaiteriez-vous travailler dans ce genre de structure ?	Oui, l’ambiance était très conviviale avec 2 réunions organisées avec 2 groupes différents chaque semaine pour partager ses avancées, ses questions et ses conseils. De plus, la recherche académique est très intéressante dans le sens où les sujets d’études sont novateurs, et que l’on peut essayer des nouvelles choses sans trop se soucier d’une obligation monétaire. Cette liberté de travail laisse plus de place à la créativité. De plus, mon équipe était essentiellement composée de bio-informaticiens, ce qui laisse une opportunité très grande à recevoir des conseils sur son travail. Mais à la fois cette équipe est très ouverte dans les séminaires et l’aide aux autres équipes de l’institut ce qui laisse place à la collaboration avec des laboratoires « paillasses » et donc ce qui permet d’avoir des dataset très intéressant pour l’analyse de données.
Ce stage vous a-t-il apporté des éléments pour préciser votre projet professionnel ?	Ce stage m’a permis de découvrir que j’aimais tout particulièrement les analyses omiques comme la transcriptomique spatiale, ainsi que le domaine du ML que je trouve passionnant et plein de potentiel. J’aimerais cependant pouvoir travailler sur des données du cerveau car les neurosciences m’ont toujours passionnées. De plus, j’aimerais aussi pouvoir comparer cette expérience à une expérience professionnelle dans une entreprise privée afin de voir mes préférences.