UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI BIOLOGIA

Corso di Laurea magistrale in Molecular Biology

TESI DI LAUREA

Computation of cancer gene expression signatures in spatial transcriptomics data

Relatore: Dott.ssa Enrica Calura

Dipartimento di Biologia

Laureando: Anna Corrà

Anno Accademico 2021/2022

Abstract

Even though tumor originates from clones of cells, it develops a substantial intratumor heterogeneity in terms of cellular morphology, gene expression, proliferative and metastatic potential. Due to this heterogeneity, diagnostic and prognostic gene expression cancer signatures often fail in the evaluation of bulk gene expression tumor profiles. Moreover, the internal organization of tumors has potential consequences on treatment response and resistance. Therefore, discerning the complexity of both composition and internal structure could provide a valid step towards the understanding of tumor biology.

Spatial transcriptomics is a new approach in the analysis of transcriptomes that allows the analysis of gene expression level in the intact tissue, maintaining the spatial information.
During my thesis project I worked in the application of cancer gene expression signatures on spatial breast cancer transcriptome data, highlighting how the resulting panel of spatially resolved cancer gene expression scores provide powerful information in tumor data interpretation.

Contents

1 Introduction

1.1 The human cancers

Cancer is a disease defined by an abnormal growth of cells caused by genetic alterations impacting the gene expressions that lead to an unbalanced condition between cell proliferation and cell death. The dysregulation of gene expression enables cancer cells to acquire, also, invasion ability, leading to the formation of metastasize in distant sites. The alteration of gene expression can occur by direct modification of DNA, such as gene mutations, translocations, amplifications, deletions, loss of heterozygosity or through mechanisms resulting in abnormal gene transcription or translation. In particular, there are some genes that have relevant importance, if mutated, in the triggering of the tumor establishment: proto-oncogenes and oncosuppressors.

The vast majority of proto-oncogenes encode for proteins that generally control cell proliferation, apoptosis or both. The products of proto-oncogenes can be many, but they are mainly transcription factors, chromatin remodelers, growth factors and growth factor receptors, signal transducers, apoptosis regulators ecc. Once alterated, the proto-oncogene becomes an oncogene that confers a growth advantage and increased survival of cells carrying such alterations .

On the other hand, oncosuppressor genes, as the name suggests, are genes involved in the opposition of those mechanisms that induce tumor generation. Tumor suppressor genes are involved in DNA damage repair, inhibition of cell division, induction of apoptosis and suppression of metastasis. Therefore, the loss of a oncosuppressive function would result in the initiation and progression of cancer. Both oncosuppressors and oncogenes have crucial roles in tumor formation, infact, most carcinomas are initiated by the loss of function of a tumor-suppressor gene, followed by alterations in oncogenes and tumor-suppressor genes.

Despite many specificities across the different types of cancers, mainly defined by the different driver genes and tissue of growth, cancers share many traits and ways of action that have been summarized in the famous reviews by Hanahan and called "hallmarks of cancers". The most updated list of all cancer hallmarks are reported in Figure 1.

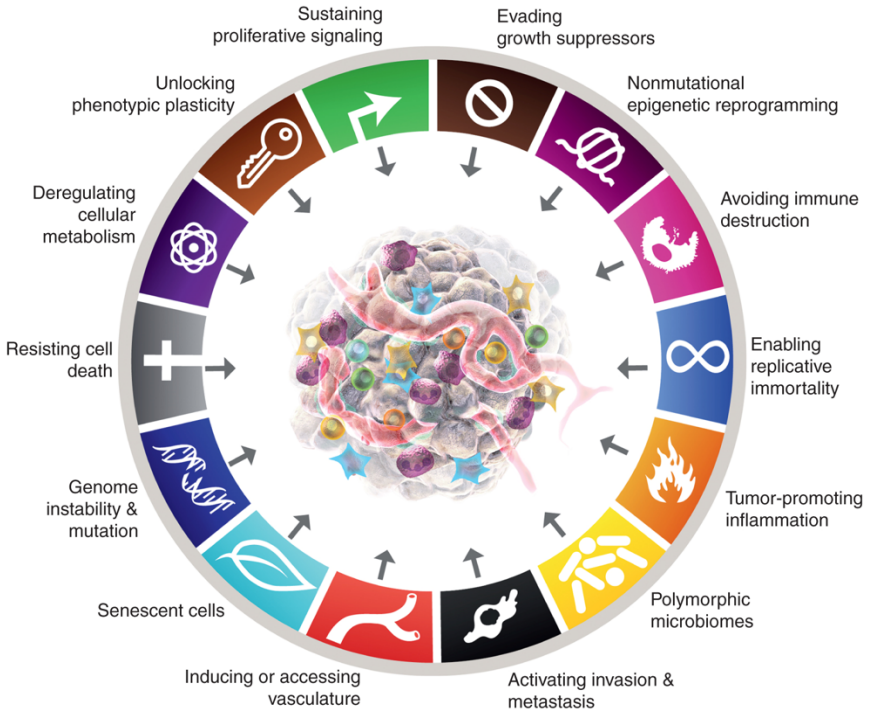


Figure 1: The hallmarks of cancer. Figure credits to Hannan, Hallmarks of Cancer: New Dimensions, *Cancer Discovery* (2022).

During the transformation and tumor progression of a normal cell toward the neoplastic state, the cell acquires several capabilities that can be categorized in: sustainment of proliferative signaling, evasion of growth suppressors, resistance to cell death, capability of replicative immortality, angiogenesis, tumor promoting inflammation, avoiding immune destruction, invasion and metastasis [1].

One of the cancer hallmarks is the establishment of a tumor microenvironment (TME) supportive of tumor traits, TME refers to the healthy cellular environment in which tumors establish and develop. The TME is determined by different components: blood vessels, immune cells, extracellular matrix (ECM), fibroblasts, lymphocytes, bone marrow-derived inflammatory cells, and signaling molecules. The TME includes, also, non-malignant cells that play a protumorigenic role in the phases of carcinogenesis.

The first component of TME are endothelial cells, which play a key role in tumor development and tumor cell protection from the immune system. Angiogenesis is a key process for the sustainment of the tumor mass and nutritional support for tumor growth. Cancer cells are able to induce the production of growth factors and stimulate new vasculature from preexisting vessels or derived from endothelial progenitor cells.

Another important component of TME are the immune cells such as granulocytes, lymphocytes, and macrophages, involved in the inflammatory reactions promoted by the tumor. The most prominent immune cell type in the TME is the macrophage that can suppress antitumor immune mechanisms and promote the escape of tumor cells into the circulatory system. Fibroblasts present in the TME are called cancer associated fibroblasts (CAFs) which, with the extracellular matrix, influence the migration of cancer cells by altering the physical properties and composition of the surrounding area of the tumor.
The acquisition of specific tumor abilities jointly with a supportive TME allows the establishment of the malignant state and the limitless growth with self-mantainment of the tumor. In this perspective, the evaluation of the presence of hallmarks in a tumor sample can be useful to characterize the neoplastic disease and, since hallmarks are defined by peculiar gene expressions, they can be studied through the definition of cancer gene expression signatures.

Cancer is a major public health problem worldwide and is the second leading cause of death, following cardiovascular disease. The survival rate of cancer patients varies abundantly between cancer types, available treatment and stage of diagnosis, fluctuating between 20% and 80%. Anyway, it remains far away from the statistics we hope for. Cancer treatments are not always effective and the onset of the proper therapeutic approach remains a challenging step. In conclusion, the overall situation highlights the incomplete understanding of the disease and the necessity of new techniques to characterize and investigate it in a deeper way.

1.2 Cancer gene expression signatures

The vast majority of tumor classificators are mainly established through an anatomo pathological evaluation, and it is based on phenotypic features such as the size of the tumor, some histological characteristics, such as the presence of tumor histotypes, the degree of spread, or the grade of nuclear atypia of cancer cells, *i.e.* the status of the cell transformation These morphological aspects, supported by clinical information, are used to classify patients and determine the most likely to benefit therapy. Unfortunately, in most cases a significant minority of patients do not respond to the treatment or show no meaningful improvements, highlighting the need for features able to capture more efficiently patient predictive specificities that are invisible to traditional classification approaches [2, 3].

Only recently, the possibility to investigate genetic aberrations of tumors brought by sequencing technologies in clinics provided additional hints for the classification of specific tumors and they are helpful in therapy choice.
It is widely recognised that new biomarkers predicting the likelihood of therapy response, or of its toxicity, are essential to allow us to better tailor treatments. Paramount is the identification of dynamic markers, in addition to predictive tools, that seek to shed light on the evolution of tumors over time and with treatment. In this scenario tumor gene expressions can further help in characterizing tumor specific ability and established or emergent therapy-relevant pathway signals.
The presence or the effects of all the cancer hallmarks described in the previous paragraph can be measured by using gene expression of cancer biopsies and through these quantitative traits many information about cancer ongoing processes can be captured.
Gene expression is a valid representation of cellular activity, while one single aberrant gene does not have enough power to define a biological state, but a set of genes can be helpful in determining cell process activities. In this context gene expression signatures are defined by a pattern of gene expressions, found unique in a specific biological scenario [2].

It has been demonstrated that in cancer gene expression signatures can be a useful classification system. In addition to increasing the understanding of the molecular mechanisms of the tumor, the signatures can be used as diagnostic, prognostic or predictive markers. Intuitively, a diagnostic signature is used in the identification of a specific clinical condition, aiming to simplify the diagnosis process. Prognostic markers help in the estimation of the most likely outcome of the cancer disease in an untreated situation. Therefore, prognostic signatures are used in the classification of patients with high risk of metastasis formation and so valids candidates for adjuvant therapy, such as chemotherapy, radiation and immunotherapy. On the other hand, predictive signatures provide information on the likely advantage gained from therapy. Such markers are useful to stratify patients and select the most proper therapeutic approach based on the benefits likely to be obtained [4].

One of the main problems in the use of omic technologies in clinics is the intra-tumor heterogeneity. In bulk transcriptome analysis all transcriptomes of all cells in the sample are analyzed together, as a consequence, specific features of different cell types are partially masked. In particular, cancer bulk transcriptome analysis results provide an estimation on the average behavior of cancer cells, losing all the cell differences that determine tumor heterogeneity. Moreover, it

has been demonstrated that these results have strong dependence on the piece of tumor sampled, suggesting that single bulk transcriptome analysis can lead to underestimation of the tumor genomics landscape [5].

Initially, cancer cells have been portrayed as a quite homogeneous cell population until the further progression of the tumor in different metastases. However, in many tumors hyperproliferation and genetic instability lead to the accumulation of multiple genetic differences that drives distinct clonal subpopulations in the same mass. Due to these clonal heterogeneity, many human tumors contain regions characterized by various degrees of differentiation, proliferation, vascularity, inflammation and invasiveness. In addition, human tumors show a repertoire of recruited apparently normal cells that enrich even more tumor complexity and intra-tumor heterogeneity since they have a different spatial location in the tissues [1].

In the end, the intra-sample tumor heterogeneity could hamper the efficacy of gene expression signatures, and in general of all the tumor biomarkers evaluated through a single biopsy. On the other hand, the study of intra-sample tumor heterogeneity of biomarkers can help to capture sample traits essential for prognostic and predictive evaluation. Nowadays, intra-sample tumor heterogeneity of gene expressions can be accessed with the state-of-the-art spatial transcriptomics technologies.

1.3 Spatial transcriptomics

Spatial transcriptomics is a technology that introduces the spatial information in transcriptome analysis. The occurrence of this new strategy allows to maintain positional information during transcriptional analysis, obtaining for one sample thousands of expression profiles spatially distributed in the tissue.
The first spatial transcriptomics technology has been proposed in 2016, in the following years some improvements in the technical approach have been implemented till the coronation of spatial transcriptomics as Nature's Method of the Year 2020.
The firsts and the currently more popular spatial technologies imply the definition of the key structure called "spot". One spot refers to the round area that is actually able to sample mRNA molecules from the tissue. These spots -sometimes called features- change in their characteristics among different technologies but, in general, they are made up by a multitude of oligonucleotides able to hybridize with the mRNA released from the tissue. Independently from the technology, spatial transcriptome profiles are always linked to the histological

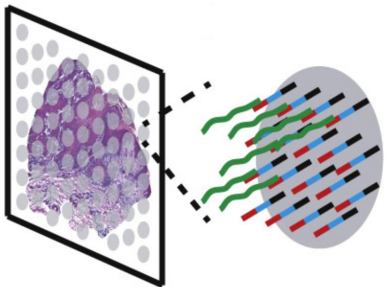


image of the tissue, which is necessary to get the morphological information and locate the several spots profiles in the tissue.

Figure 2: Spatial transcriptomics capture spot. Figure credits to Ji, Multimodal Analysis of Composition and Spatial Architecture in Human Squamous Cell Carcinoma, *Cell* (2020) [6].

Resolution analysis varies based on the technology applied, from 20 cells (belonging to the same "spot") up to a fine subcellular resolution.
The integration of spatial information brings multiple advantages in gene expression analysis. First, maintaining spatial information, clearly, gives the possibility to identify the distribution of gene expression together with the cell types inside a tissue and keep track of their relative position. Second, spatial information can also be used as a criteria to share information between adjacent spots. For example, depending on the expression profile detected, each cell (or spot) can be categorized and those with similar profiles are assigned to the same group. Due to technical limitation or simply by experimental imprecision there can be cells (or spots) without clear identity or behavior. Assuming that it is likely that close cells belong to the same group type, the spatial information can be exploited to infer information based on the neighborhood.

Moreover, it is known that the internal structure of tissue matters, such as the correlation between the type of immune cells and their organization in the tumor mass and TME and the disease progression [7, 8]. Thanks to spatial technologies we can understand the internal organization of a tissue, thus in tumor sample analysis we can use specific spatial information as a criteria to stratify patients and move towards a proper therapeutic approach.
Currently, there are two main categories of spatially resolved technologies: Next-Generation Sequencing-based and images-based approaches. The first one's process, before NGS, encodes spatial information in the transcriptome, instead the second one exploits the hybridization and imaging of probes directly onto the tissue [9].
Here below a review of the main spatial transcriptomics technologies available.

1.3.1 10x Genomics Visium

10x Genomics Visium is by far the most known and common technology for spatial transcriptomics. A *Visium Spatial Gene Expression Slide* has multiple

capture areas, each one of 6.5 x 6.5 mm^2, covered by 5000 spots. Each spot has a diameter of 55 µm and the distance center-to-center between them is 100 µm. A single Visium spot is made by millions of capture probes, each composed by sequencing primers, spot-unique spatial barcode, UMI and a poly-dT tail. Depending on the size of capture spots, 10x Visium is not a single cell resolution technology: according to the type of tissue each spot covers 5-20 cells.
This technology can work both with fresh frozen and Formalin-Fixed Paraffin-Embedded tissue samples. FFPE is a common type of preservation of specimens in which the tissue is first fixed in formalin and then embedded in a paraffin wax block. This kind of conservation enables to preserve the structure of the tissue and makes it easy to cut in slices the sample. A 10 µm thick slice of the tissue is obtained by cryosection (or though microtoming FFPE samples) which is then laid on the capture area and then permeabilized to allow the release of the mRNA, which is retained on the slide by the poly-dT binding tail. The following step is the extension reaction to generate a barcoded sequencing-ready library. The subsequent sequencing reveals the transcriptome content and its location onto the tissue, codified by the unambiguous barcode.

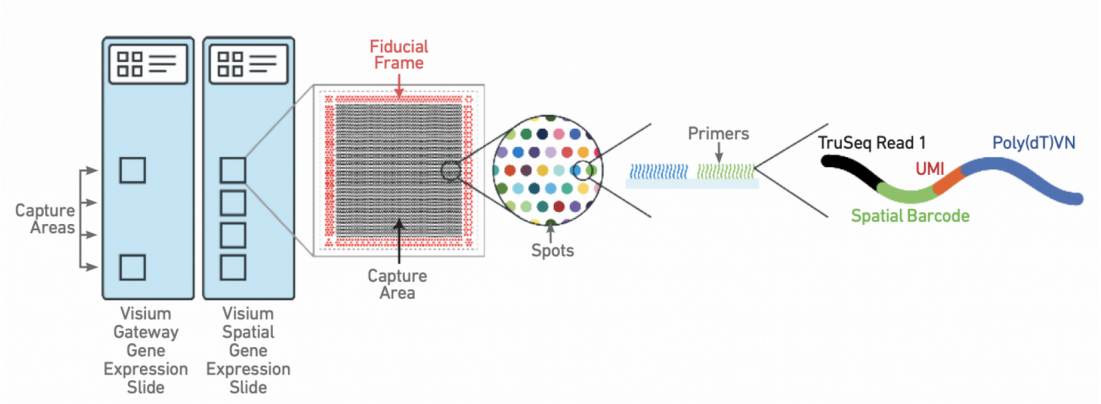


Figure 3: 10x Genomics Visium Spatial capture slides. Figure from "Visium Spatial Gene Expression."10xGenomics.(https://www.10xgenomics.com/products/spatial-gene-expression).

1.3.2 Slide Seq

Several new methods and strategies to collect mRNAs from tissue have been proposed. Slide Seq (2019) is another technology for spatial high-resolution genome-wide expression analysis. This approach takes inspiration from the Drop-seq method for single cell RNA-seq, using similar DNA barcoded microparticles.
Slide seq entails the formation of a monolayer of beads, named "puck", onto a rubber-coated glass coverslip. One puck has a diameter of 3 mm and consists of

roughly 70,000 barcoded beads. At first, the beads' barcode is uniquely determined via SOLiD sequencing-by-ligation chemistry, in order to match each unique barcode with a spatial location. Then, a 10 µm thick slice of fresh-frozen tissue is laid on the puck, the tissue is permeabilized and the mRNAs released are captured by the beads for preparation of 3'-end barcoded libraries.
Since the diameter of the beads is 10 µm, the resolution of this technique is comparable to the sizes of individual cells. The authors subsequently presented Slide seq V2 (2021), which is an improvement of Slide Seq in terms of sensitivity, maintaining the same technical approach [10].

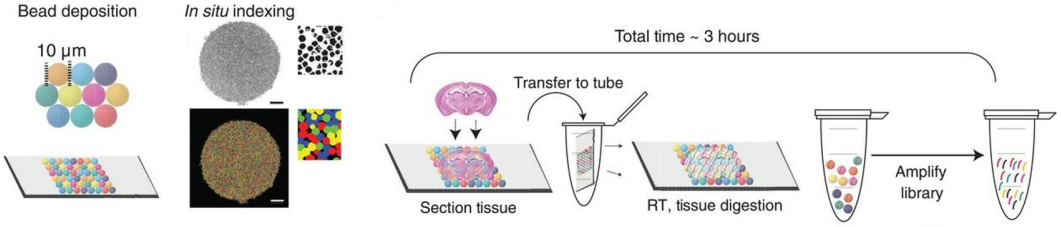


Figure 4: Schematic array generation and sample preparation procedure developed for Slide-seq. Figure credits to Rodriques, Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution, *Science* (2019) [10].

1.3.3 HSDT

High-definition spatial transcriptomics (2019) is another technique for spatial tissue profiling, providing a further improvement in resolution. HSDT, similar to Slide-seq, involves 2 µm diameter uniquely barcoded beads which are deposited onto an array of hexagonal wells. The array has a dimension 5.7 mm x 2.4 mm (13.7 mm^2) in which there are more than 1.4 million wells with a 2.05 µm diameter. To ensure that all hexagonal wells will contain one bead, this technology reaches the production of almost 3 millions of beads.
After the deposition and the decoding of beads spatial location, a frozen tissue section is placed on the array of beads. As in the previous technologies, the tissue is permeabilized to capture the RNAs for the subsequent transcriptome analysis [11].

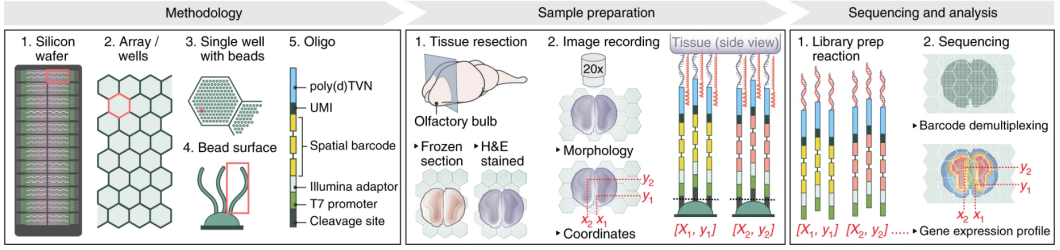


Figure 5: Schematic workflow of HSDT technology. Figure credits to Vickovic, High-definition spatial transcriptomics for in situ tissue profiling, *Nature Methods*, (2019) [11].

1.3.4 DBiT-seq

Another proposal applying a substantially different strategy in the application of barcodes is Deterministic Barcoding in Tissue sequencing (2020). DBiT-seq is a spatial omics sequencing technique able to map both mRNAs and proteins onto tissue slices. Rather than removing mRNA molecules from the tissue to label them with a spatial barcode, DBiT-seq links barcodes to molecules directly in tissue, avoiding potential lateral diffusion during the releasing of mRNAs.
This approach entails microfluidic channels to deliver two sets of barcodes on the tissue: the first set defines parallel stripes of barcodes, the second set defines orthogonal stripes, compared to the first one, creating an intersecting grid of spatial barcodes combination. The DBiT-seq microfluidic device has 50 parallel channels with a width down to 10 µm, which defines the dimension of the pixels of the technology, in which authors declared to reach the detection, on average, of 2000 genes. Thanks to the dimension of the microfluidic channels that defines the barcoding, this technology can be considered to have a single cell resolution [12].

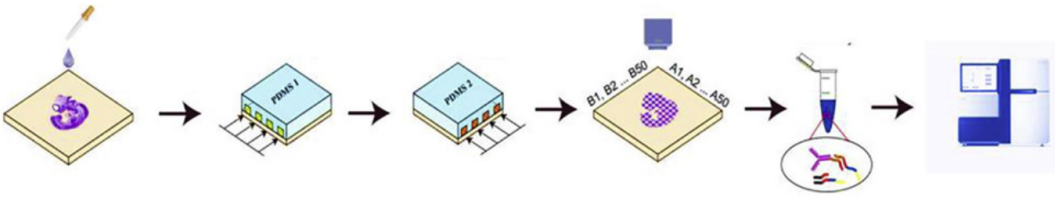


Figure 6: Schematic workflow of DBiT-seq. Figure credits to Liu, High-Spatial-Resolution Multi-Omics Sequencing via Deterministic Barcoding in Tissue, *Cell*, (2020) [12].

1.3.5 PIXEL seq

Polony (a contraction of "polymerase colony", also called a DNA cluster) **i**nde**xe**d **l**ibrary-sequencing (PIXEL seq) spatial transcriptomics is another technology that aims to improve the spatial barcoding efficiency. PIXEL seq entails the generation of "continuous" polony oligos arrayed across a gel surface: templates containing reverse transcription primers and spatial index are seeded on the gel surface and then amplified to dsDNA polonies. Then follows the digestion of the dsDNA which exposes the RT primers.
Differently from the previous technologies, PIXEL seq uses a crosslinked PAA gel (customizable size e.g. 6 × 30 mm^2) without pre-defined feature boundaries, such as beads or spots. Each feature is defined by spatial index polony sequencing by sequencing-by-synthesis chemistry, obtaining features with center-to-center distance of 1 µm. PIXEL seq is able to capture more than 1000 unique molecular identifiers in 10 μm^2 area, attaining ≤ 1 µm resolution. 90% of the overall gel area

is covered by barcoded oligos and compared to solid surface, PAA gel allows 10 to 30-fold higher oligo capturing efficiency [13].

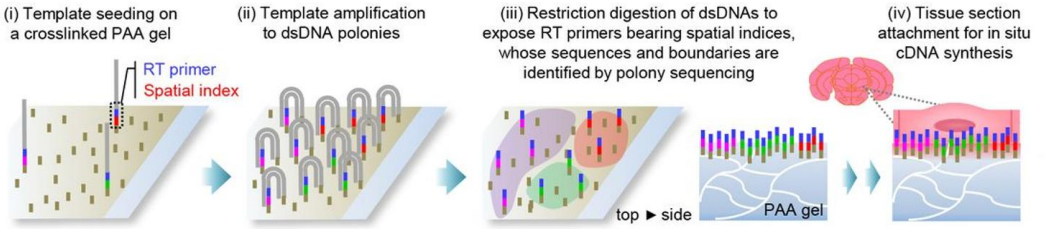


Figure 7: Scheme of PIXEL-seq-based spatial transcriptome analysis. Figure credits to Fu, Continuous Polony Gels for Tissue Mapping with High Resolution and RNA Capture Efficiency, *bioRxiv*, (2021) [13].

1.3.6 Seq Scope

Another step towards high resolution is made by Seq Scope technology (2021): 0.6 µm average center-to-center distance between features. Seq scope is based on illumina amplification of spatial barcode RNA capture molecules on a solid support, creating up to 150 clusters in 100 μm^2. The procedure consists of two sequencing steps: the first sequencing generates a spatial map of barcodes each one associated to its XY coordinates; the second one is performed on the captured mRNA molecules, after the laying of the tissue slice on the solid support and the permeabilization of it [14].

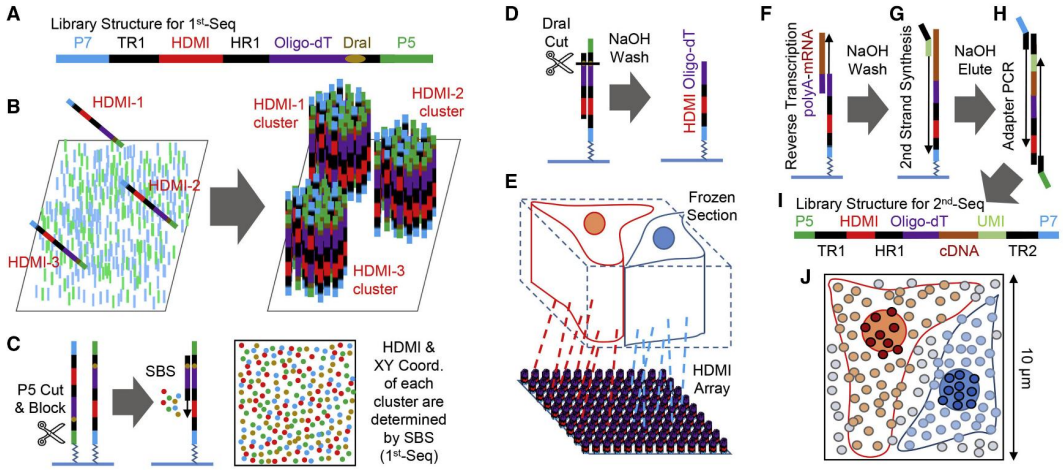


Figure 8: Seq scope technology overview. A. Schematic representation of the HDMI-oligo library structure for 1st-Seq. B. Solid-phase amplification of different HDMI-oligo molecules on the flow cell surface. C. Illumina sequencing by synthesis determines the HDMI sequence and XY coordinates of each cluster. D. Then, HDMI oligonucleotide clusters are modified to expose oligo-dT, the RNA-capture domain. E. HDMI-array captures RNA released from the overlying frozen section. F. Then, cDNA footprint is generated by reverse transcription. G. After that, secondary strand is synthesized using random priming method. H-I. Finally, adaptor PCR generates the sequencing library for

2nd-Seq, where paired-end sequencing reveals cDNA sequence and its matching HDMI barcode. J. HDMI-array contains up to 150 HDMI clusters in a 100 μm^2 area. Figure credits to Cho, Microscopic examination of spatial transcriptome using Seq-Scope, *Cell* (2021) [14].

1.3.7 Stereo Seq

The currently highest resolution spatially resolved transcriptomic technology is Stereo Seq (Spatio-Temporal Enhanced REsolution Omics-sequencing, 2021): it relies on DNA nanoball (DNB) containing random barcode sequences, placed 500/715 nm far from each other. The support surface of Stereo seq technology is a modified silicon chip photolithographically etched with a grid of 220 nm diameter spots, where DNB are deposited. Each DNB, labeled with random barcode, is generated by rolling circle amplification and allows a spatial barcode pool size of 4^{25}, larger than eads based technologies.
At first the array is microphotographed, and sequenced to link each coordinated identity (CID) to its spatial location. Then, molecular identifiers (MID) and polyT oligonucleotides are ligated to each DNB. This strategy allows the generation of 50, 100, 200 mm^2 containing barcoded spots at higher density than other previous methods. Notably, the author developed DNB patterned array chips till the dimension of 42.25 cm^2 for potential application to a whole section of the human brain [15].

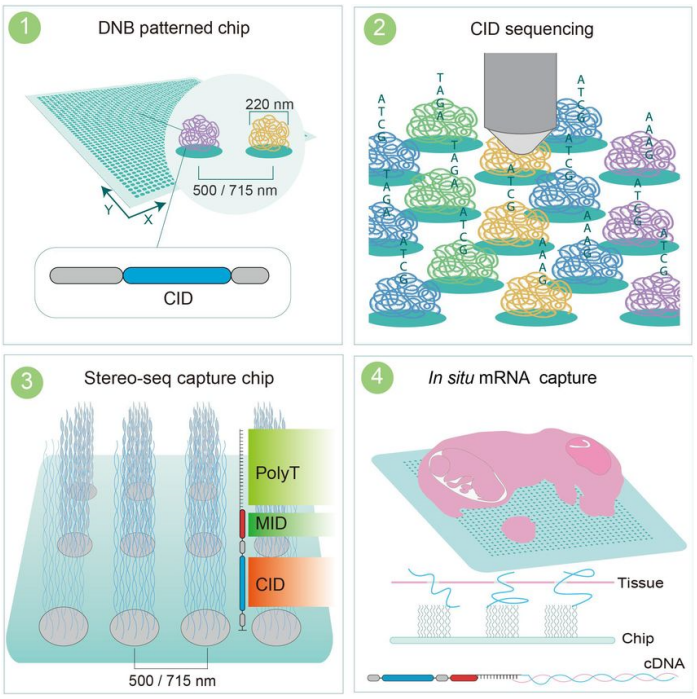


Figure 9: Schematic representation of the Stereo-seq procedure. Figure credits to Chen, Large field of view-spatially resolved transcriptomics at nanoscale resolution, *bioRxiv*, (2021) [15].

1.3.8 GeoMx

Lastly, as 10x Genomics, also Nanostring have proposed its spatial technology: the GeoMx Digital Spatial Profiler (2020). It combines both image analysis for morphological delineation and the application of NGS sequencing for the

decoding of barcodes and transcripts identification. The first step is the design of specific complementary target sequences linked to a barcode through a photocleavable linker; it follows the incubation with a panel of fluorescent antibodies and the complementary probe sequences. Fluorescent tagged antibodies, that determine the morphology of the tissue, are used to select the area of interest on which UV light is applied inducing the releasing of barcodes. A microcapillary aspirates those oligos and NanoString nCounter or Illumina NGS is used to determine and quantify them. The GeoMx probes panel allows the detection of up to 18.000 genes simultaneously and the selected area of interest can go down to the resolution of 10 µm.
GeoMx also enables the detection of proteins with the same procedure using, instead of complementary sequences, antibodies conjugated with the photocleavable linker and identificative barcode [16]. Overall, Nanostring spatial technology recalls the Laser Capture Microdissection approach, in which the selected area of interest is physically removed using laser from the tissue for subsequent sequencing analysis.

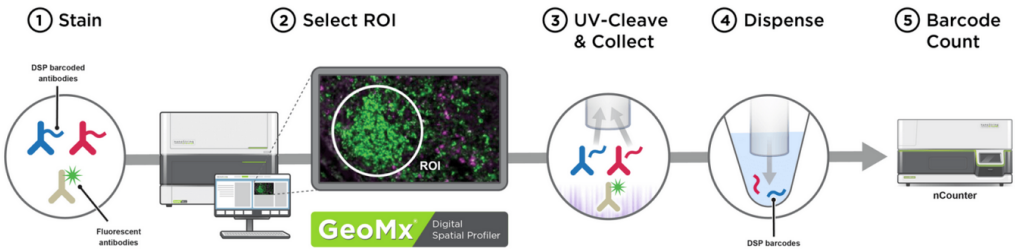


Figure 10: GeoMx workflow. Figure credits to Bergholtz, Best Practices for Spatial Profiling for Breast Cancer Research with the GeoMx® Digital Spatial Profiler, *Cancers* (2021) [16].

Merfish and seqFISH belong to a different category of spatially resolved technologies: they rely on single molecule in situ imaging methods using targeted combinatorial FISH labeling.

1.3.9 MERFISH

MERFISH (2015) stands for multiplexed error-robust FISH, it is a technique that allows the decoding of 100 up to 1000 targeted RNA species, with subcellular resolution. This approach identifies RNA molecules thanks to an error-robust encoding scheme of in situ hybridization probes. Each target transcript is associated with one specific binary code for its identification. That binary code comes from the presence or absences of fluorescence signals detected during the hybridization of design probes. Sequential images are taken and, in parallel,

multiple fluorescent signals are collected, which are subsequently translated in 0 and 1. The error-robust encoding scheme of hybridization allows to recognise possible missed fluorescence signals or hybridization, keeping the correct identification of transcripts [17].

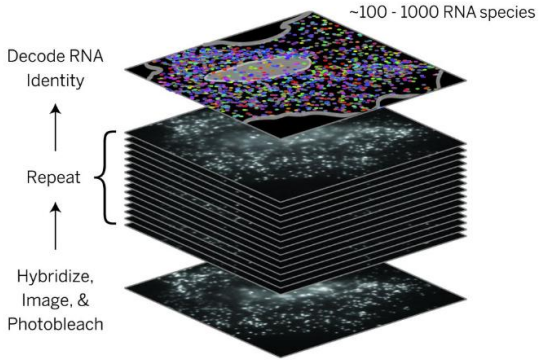


Figure 10: MERFISH workflow. Figure credits to Chen, Spatially resolved, highly multiplexed RNA profiling in single cells, *Science*, (2015) [17].

1.3.10 seqFISH+

seqFISH+ (2019) works with the same approach: the identification of RNAs is based on sequential fluorescence in situ hybridization and decoding of images. seqFISH+ improved the encoding strategy using primary target genes probes and readout probes defining an encoding system of 60 pseudocolors. The primary gene probe core is composed of a complementary sequence to the target gene, while the two tails remain howerhanging. Each tail has two complementary regions to the readout probes, so it has the possibility to hybridize with 4 different readout probes, which are linked to a specific fluorescent color. After 20 rounds of hybridizations and readouts, the combination of colors obtained is decoded, allowing the detection of up to 24000 genes [18].

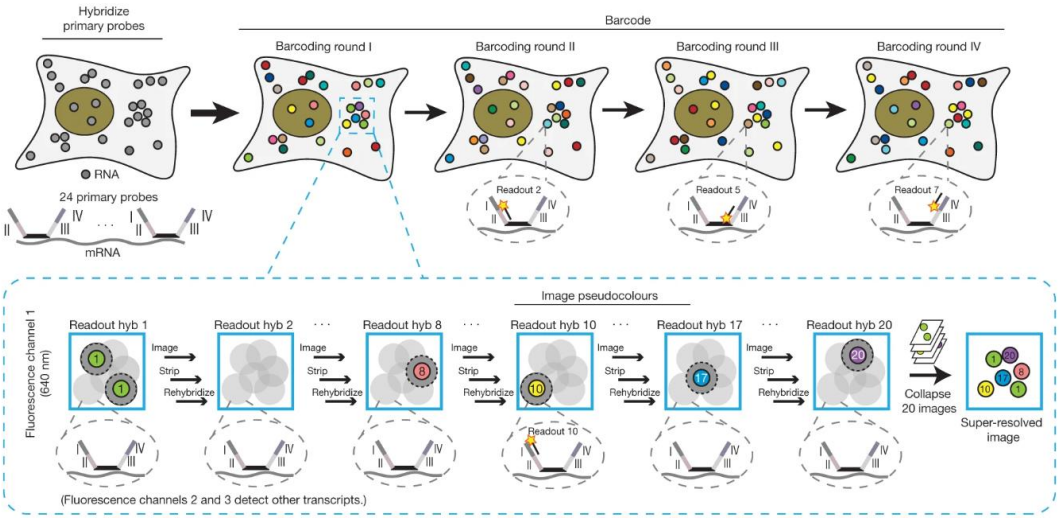


Figure 12: seqFISH+ schematic pseudocolor coding. Linus Eng, Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+, *Nature* (2019) [18].

1.3.11 STARmap

Another type of image-based approach is Spatially-resolved Transcript Amplicon Readout mapping (STARmap, 2018), which involves rolling amplification and in situ sequencing. STARmap starts with the hybridization of SNAIL probes to target mRNAs, ligation and amplification of it in the tissue creating a cDNA SNAIL amplicon.

STARmap aim's is to give a portrait of the gene expression spatially distributed in the threedimenstioanl space, which is obtained by the embedding of the cDNA amplicon in a tissue hydrogel setting. Amine-modified nucleotides are spiked into the rolling-ciciel ampicon reaction and copolymerized with acrylamide monomers to form a stable crosslinked hydrogel-DNA amplicon network. Each SNAIL probe contains a 5 base gene-specific identifier which, once amplified, can be readout by fluorescence in situ sequencing. STARmap has single cell resolution and is able to identify and map up to 1000 genes in each section, over six imaging circles [19].

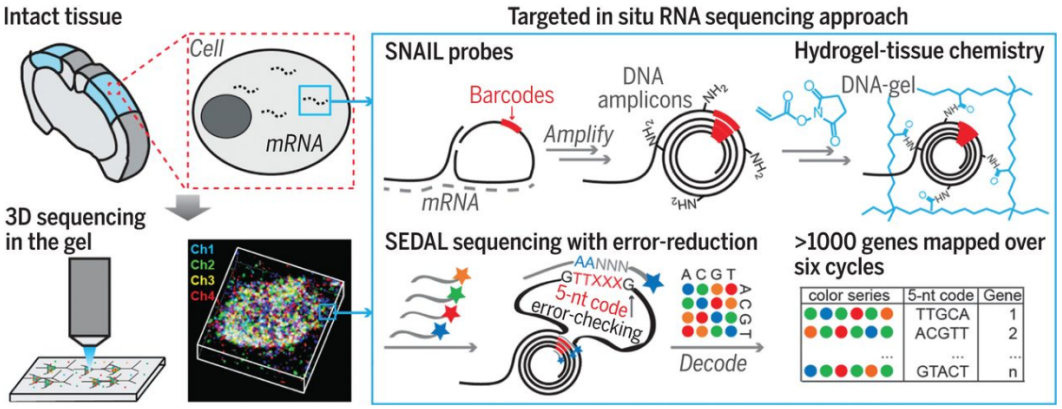


Figure 13: Schematic representation of STARmap technology approach. Figure credits to Wang, Three-dimensional intact-tissue sequencing of single-cell transcriptional states, *Science* (2018) [19].

1.4 Spatial transcriptomics data general properties

Due to its quasi-single cell resolution, spatial transcriptomics data presents common characteristics with single cell transcriptomics data, both in terms of magnitude and distribution of count values.

The features of single cell data have been well documented in the work of Lahnemann et al. [20] in "*Eleven grand challenges in single-cell data science*", in which the authors outlined several prickliness of single cell data. Based on the similarity between spatial and single cell, those considerations can be equivalently applied on spatial transcriptomics data.

Spatial data are characterized by a high level of sparsity defined by a large fraction of zeros in the count matrix. Those zero observations are defined as "dropouts" that usually conflate two distinct types of zero values: true and artificial zeros. True zeros are those observations that indicate biologically true absence of expression, while artificial zeros are due to genes that are expressed but not detected by the sequencing technology. Those artificial zeros are attributed to technical limitations and can be either systematic (e.g. specific mRNA degradation) or can occur by chance (e.g. barely expressed transcripts that sometimes will be detected and sometimes not). Beyond biological variation in the number of unexpressed genes, these proportions of artificial zeros considerably contribute to the level of sparsity of the data [20].
In addition, spatial expression count values represent the number of expression counts collected in one spot covering a few cells, as a consequence the number of detected genes and counts will be restricted. The low count values along with the large fraction of zeros give spatial transcriptomics data highly susceptibility to technical noise.

Spatial information gives the possibility to handle sparsity and denoising spatial data by the construction of a smoothing method based on neighbors information. Smoothing methods mainly allow the sharing of gene counts between close spots resulting in a reduction of zeros observation and in an increase of counts expression values.

1.5 Computational analysis of spatial transcriptomics data

From the release of spatial transcriptomics technology in 2018, several specific spatial transcriptome analyses have been developed. At the beginning, the type of analysis was very similar to the approach used with single-cell data, with the additional spatial visualization of the results on a tissue slice. More recently have been developed different packages that exploit the spatial information more fruitfully, including it in smoothing methods and different steps during the analysis. Below, it's reported an overview of how some of those packages propose to integrate spatial information in the analysis and how to overcome noise and sparsity of data.

1.5.1 stLearn

stLearn is a python package developed for the analysis of spatial transcriptomics data. This software, as inputs, needs the gene expression count matrix distributed in the tissue slice, the spatial coordinates of those counts and the image of the tissue. *stLearn* allows to integrate gene expression, spatial location

and tissue morphology in one comprehensive analysis that, as well as detecting cell types, include the possibility to find regions with high cell-to-cell interactions and to reconstruct cell trajectories.

The strength of *stLearn* is the effective usage of spatial information in the analysis of expression data by the implementation of Spatial Morphological gene Expression normalization (SME normalization). It's an within-tissue normalization step, performed upstream of all the analysis, based on morphological similarity and neighborhood smoothing. Spatial information implies the definition of a neighbor area for each spot (or cell), which includes the set of those spots (or cells) within a given radius *d.* Morphological similarity is a value considered as "distance" between neighbor spots proportional to the similarity between the morphology of their underlying tissue. The features that define the morphology of the H&E images of the tissue are extracted by a convolutional neural network (CNN) model, widely used for image classification. Based on the assumption that close and morphologically similar spots are more likely to cover the same cell types and so to have similar gene expression profiles, *stLearn* applies on each spot the neighbor smoothing. The neighbor smoothing is a gene expression values adjustment in which the expression value of one specific gene in a central spot is summed up to the mean of the expression values of that gene in the neighbor spots, weighted on the morphological similarity.

As previously mentioned, spatial gene expression data is characterized by a large amount of zero counts, in this way a zero-expression value will be maintained in a spot only if all the surrounding spots have zero counts too. By applying the SME normalization, *stLearn* allows the sharing of information between spots, leveraging the spatial knowledge that only spatial technologies are able to provide.

stLearn can be applied on different type of spatial input data such as 10x Visium (and older ST), Slide-seq, MERFISH and seqFISH, but the SME normalization does not apply to image-based technology, which have not a predefined subsetting of the tissue slice [21].

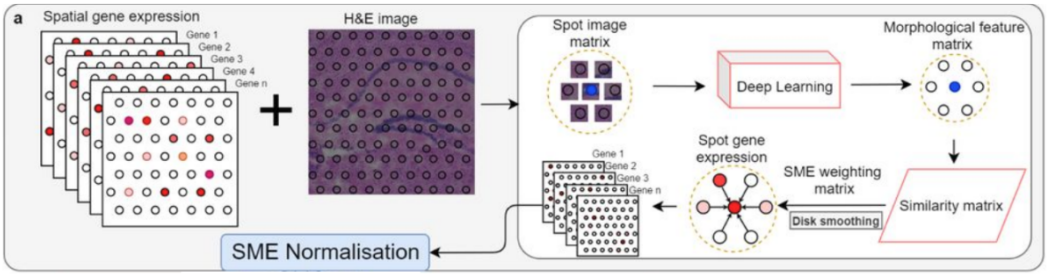


Figure 14: *stLearn* integration of H&E image through deep learning approach. Figure credits to Pham, *stLearn*: integrating spatial location, tissue morphology and gene

expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues, *bioRxiv*, (2020) [21].

Currently, there are other packages that are designed for spatial data that offer the possibility to exploit spatial information in the analysis. However, they propose approaches that use the spatial information in a more marginal way than the stLeran approach.

1.5.2 Giotto

Giotto is an R package drawn for spatial transcriptomics data analysis and visualization that allows the characterization of tissue composition, spatial expression pattern and cellular interaction. *Giotto* works with different spatial input data coming from different spatial technologies and single-cell RNAseq data can be integrated for spatial cell-type enriched analysis.
This package uses spatial location, clearly, in the visualization of the results projected on the tissue section and integrates it in the analysis with the implementation of two types of structures: grid and network. For those spatial technologies with single molecule resolution (MERFISH, seqFISH+), *Giotto* proposes a spatial grid for pooling together those expression values located close to each other. With this approach the gene expression level of cells within each box grid are averaged for further analysis. Alternatively, for single cell resolution data, *Giotto* builds a neighborhood network, that is a graphical representation of neighbor cells linked one to the other through edges. These two structures are used in the detection of spatial coherent gene expression, proximal cell types, staptial co-expression pattern and discrete domains [22].

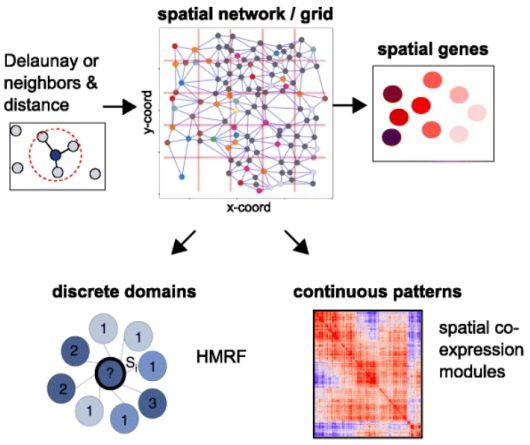


Figure 15: Schematic representation of *Giotto* spatial analysis. Figure credits to Dries, *Giotto*: a toolbox for integrative analysis and visualization of spatial expression data, *Genome Biology*, (2021) [22].

1.5.3 spaGCN

Also *spaGCN* is a package that integrates gene expression, spatial location and histology. *SpaGCN* is a python package developed for the analysis of spatial

transcriptomic data, whose focus is to identify spatial domain and spatially variable genes. It is applicable to both in-situ transcriptomics with single-cell resolution and spatial barcoding based transcriptomics data.

First of all, *spaGCN* integrates spatial location and histology information building an undirected weighted graph to represent the relationship between all spots. The edge weight between two spots is defined by the euclidean distance between them, calculated on the two spatial coordinates (x, y) and a third dimension (z). The third dimension z is given by the information extracted from the histology image under the spot: *spaGCN* calculates z based on the mean and variance of the colors of the spots under consideration. Larger variance will be translated with higher z values. In this way the undirected weighted graph represents the spatial dependency of the data, deriving both form spatial location and histology features.

Then, in the process of the identification of spatial domain, this software utilizes graph convolutional layers to aggregate gene expression information from neighbor sports, according to the specific edge weight. As well as *stLearn*, *spaGCN* in this way implements the sharing of gene expression information between spots based on spatial location and weighted on the morphology of the tissue [23].

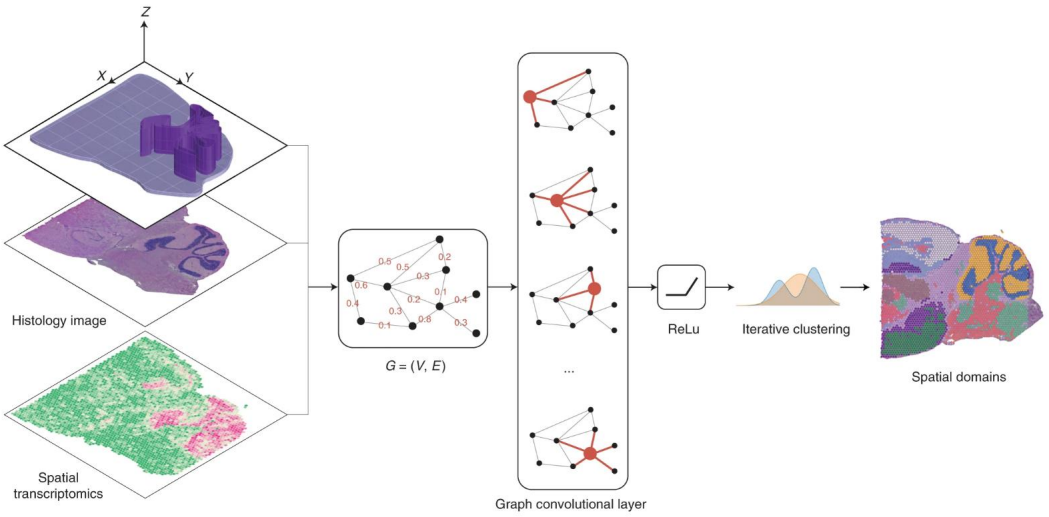


Figure 16: Overview of *spaGCN* workflow. Figure credits to Hu, *SpaGCN*: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network, *Nature Methods*, (2021) [23].

1.5.4 BayesSpace

BayesSpace is an R package developed for increasing quality and resolution in the analysis of spatial transcriptomics data. *BayesSpace* implements a complex Bayesian statistical model in which it leverages neighbor information to increase

the resolution up to subspot level. Each spot is translated into an hexagonal spot, thus an hexagonal grid is defined that naturally determines the neighborhood structure. The approach entails the subdivision of each hexagonal spot into six sub-spot whose gene expression level is estimated through Markov chain Monte Carlo. The prediction of each sub-spots expression is based on the estimated expression level of the neighbor subplots, maintaining fixed the observed gene expression of the complete spot.
Also the clustering method proposed by the authors highlights the integration of spatial information: neighboring spots are encouraged to belong to the same cluster.
With this approach the authors demonstrate that *BayesSpace* outperforms the results of other spatial analysis methods in the clustering and recognition of the tissue layers, being extremely close to the manual annotation [24].

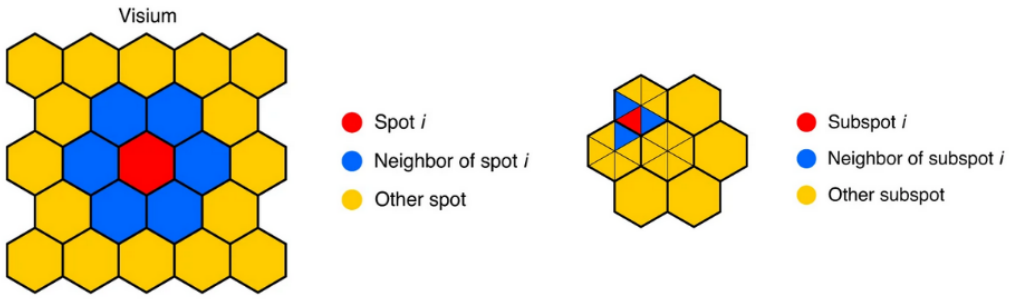


Figure 17: Representation of the Bayespace subsection of 10x Visium spots grid. Figure credits to Zhao, *BayesSpace* enables the robust characterization of spatial gene expression architecture in tissue sections at increased resolution, *bioRxiv*, (2020) [24].

1.5.5 STUtility

STUtility is an R package developed for 10x Visium data analysis and visualization. Differently from previous packages, it allows the user to visualize consecutive stacked images to get an holistic 3D view of the tissue. *STUtility* uses the spatial information implementing the spatial autocorrelation: it's an analysis for the identification of genes with clear spatial expression patterns. First, a connection network for each capture-spot is created, which links each spot with its neighbors, defined as the set of spots within a radius of 150 µm. This network is used to compute the spatial lag vector for each gene, which is defined by the summed expression of that gene across the neighbors spots to the spot under consideration. *STUtility* calculates the Pearson correlation between the lag vector (of a specific gene) and the expression count vector (for the specific gene in the central spot), which then is used to determine the spatial correlation across the whole tissue [25].

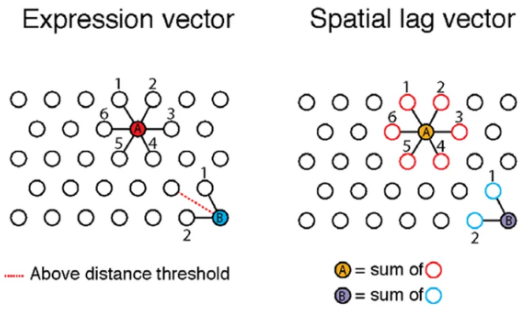


Figure 18: *STUtility* identification of neighbor spots in the computation of spatial correlation. Figure credits to Bergenstrahle, Seamless integration of image and molecular analysis for spatial transcriptomics workflows, *BMC Genomics*, (2020) [25].

Even though this is far from the approach of integrating spatial location and morphological features, this can represent a good starting point for the development of a more complex approach or usage of the spatial information.

1.5.6 Sliding window approach

Another potential inspiring approach can be the implementation of the *sliding window* by Andrew L. Ji *et al.* for spatial correlation analysis. In this study spatial correlation indicates if, defined one central spot, the expression of a gene in the surrounding spots is correlated to the expression level of the "anchoring" gene in the central spot. The authors reasoned that simple correlation between genes across spots could miss this type of expression correlation, between the set of adjacent spots. Therefore, they calculated average gene expression considering one central spot and its surrounding ones, and moving one spot further across the whole tissue. With this *sliding window* approach they generate a matrix of average gene expression that can be correlated with any "anchoring" gene of interest [6].

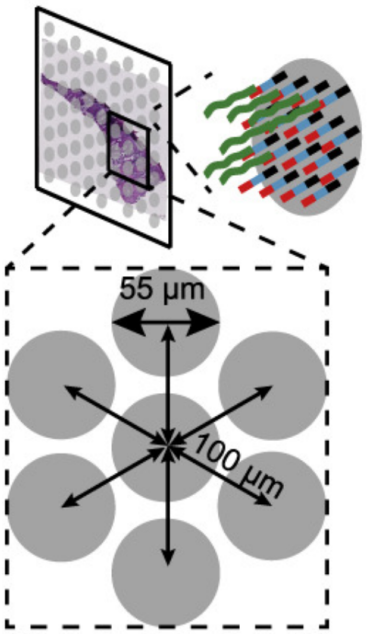


Figure 19: Representation of the sliding window concept across the slide spots. Figure credits to Ji, Multimodal Analysis of Composition and Spatial Architecture in Human Squamous Cell Carcinoma, *Cell* (2020) [6].

2 The Aim and Rationale of My Project Thesis

The project of my thesis was focused on developing a pipeline dedicated to explore the possible application of the existing cancer gene expression signatures to spatial transcriptomic data. As far as we know at the time of this writing, this is the first time that someone is trying to address this task. We retain that the application of cancer expression signatures in a spatial context could be an advantageous tool to better investigate tumor samples in clinics, where cancer gene expression signatures could be used to monitor treatment efficacy and patient prognosis. Additionally, this practice could provide useful information on the intra-sample heterogeneity of signatures, which is one of the highest causes of failure of gene expression biomarkers when translated into clinical practice.

To achieve this task, first I explored, studied and collected the latest spatial technologies proposed in literature, as well as the spatial analysis packages, especially designed to address the sparsity problem of these data. Therefore, I obtained an overview on the current spatial transcriptome data and analyses. Then since the focus of my project was on cancer, I collected all the publicly available spatial transcriptomic data related to human cancers, which were mainly studied with the 10x Genomics Visium technology, and I analyzed all of them. The analysis of my thesis includes the application of a standard pipeline for the analysis of spatial data and the calculation of multiple cancer gene expression signatures on these samples. I provided in the following chapters the results of my analysis applied to a 10x Visium data of a human breast ductal carcinoma sample taken as a case study. The sample was analyzed using more than 40 gene expression signatures, some of them provided interesting hints for a deeper understanding of the tumor biopsy.

3 Materials and Methods

3.1 The 10x Genomics Visium data of a human breast ductal carcinoma sample

For demonstration purposes, the 10x Genomics provides a large set of publicly available spatial transcriptomic data. The 10x Visum dataset contains data obtained by the analysis of both healthy and cancerous tissues, from mouse and human patients. For each sample 10x Genomics provides the raw input files: the FASTQs, the histologic image (TIFF format) and the coordinate and information about the spots. 10x Genomics also provide a standard analysis which include the alignment of the sequenced read to the gene set (the BAM file), and multiple pre-molecule read information, the filtered and raw count matrix, and also some results of a basic analysis performed with their proprietary software called Space Ranger - a set of analysis pipelines that exclusively process Visium Spatial Gene Expression data.
Even if I explored the vast majority of the available sample, for demonstration purposes all the analysis presented in this thesis was carried out on the spatial transcriptomic experiment of the sample called "Human Breast Cancer: Ductal Carcinoma In Situ, Invasive Carcinoma (FFPE)". This data comes from a 73 years old asian woman following the Visium spatial gene expression protocol for Formalin-Fixed Paraffin-Embedded specimens, which gives the possibility to perform spatial Visisum analysis not only to fresh frozen samples.

The sequencing of these data is obtained using Illumina NovaSeq sequencer with a depth of 32,524 reads per spot, for a total amount of almost 82 Million of Reads. 2,518 out of 5,000 spots have been identified under the tissue and the median number of genes detected per spot are 5,244 [26].

3.2 Software and tools used by the analysis procedure

The data previously described has been downloaded from the 10x website (https://www.10xgenomics.com/) where it has been processed using Space Ranger software version 1.3.0. Space Ranger, the proprietary software of 10x Genomics, is a set of analysis pipelines that process Visium Spatial Gene Expression data. It has been used for the demultiplexing of the Visium-prepared raw base call (BCL) files generated by Illumina sequencers into FASTQ files. This task is accomplished by a wrapper around Illumina's bcl2fastq with additional features that are specific to 10x Genomics libraries and a simplified sample sheet format. Finally, Space Ranger takes a microscope slide image and FASTQ files and

performs alignment, tissue detection, fiducial detection, and barcode/UMI counting. The pipeline uses the Visium spatial barcodes to generate feature-barcode matrices, determine clusters, and provide the gene expression matrix.

Then, following the anatomopathologist instructions about the analysis of the histologic image, the manual annotation of the spots of the spatial data have been provided using the Loupe Browser interface. Loupe Browser (Version 3.0, 10x Genomics) is a desktop application that provides programming free interactive visualization functionality to analyze data from different 10x Genomics solutions. Loupe Browser allows the user to easily interrogate different views of the Visium Spatial Gene Expression data. In addition to the downstream visualization and annotation capabilities, Loupe Browser offers support for manual alignment of fiducial frame and tissue selection upstream of running Space Ranger pipeline.

R (Version 4.2.1) is a language used for computing and graphics. I performed all the analysis supported by RStudio integrated development environment (Version 2022.2.2, build 485). The vast majority of R packages used by the analysis procedure belong to the Bioconductor platform (Version 3.15). Bioconductor is an open source and open development software repository that hosts a wide selection of analysis tools for computational biology and bioinformatics. The compatibility of my analysis with the Bioconductor environment is strategic for its usability and distribution. Thus, in my analysis I used the class of data and utilities provided by the R package *SpatialExperiment* (Version 1.6.1) to efficiently work with spatial transcriptomics data in the R Bioconductor environment. The package provides an object of class S4 to store expression matrix, sample and genes annotations, spatial coordinates, images, and image metadata. *SpatialExperiment* also includes a specialized constructor function that I used to load the data from the 10x Genomics Visium platform into a *SpatialExperiment* object [27].
The first step in the analysis of the breast cancer spatial transcriptomics data has been the normalization of the raw counts. The normalization applied on the spatial data, based on the previously mentioned similarities to single cell transcriptomic data, is the one implemented in the single-cell designed package scater (Version 1.24.0). The R package scater offers a collection of analysis tools for single-cell RNA-seq gene expression data suitable also with the *SpatialExperiment* object [28].

The visualization of the spot annotation and the signature score distribution among spots have been performed using *ggspavis* (Version 1.2.0). *ggspavis* is an R package designed for the visualization of spatially resolved transcriptomics data, enabling in particular the display of 10x Visium spots arrangement [29].

Finally, I used *signifinder* (Github), an under development R package that allows me to compute a collection of cancer gene expression signatures on transcriptomics data. *signifinder* is an R package currently present and downloadable from Github. Github is an internet hosting service for software development and version control.
In my work I used and I had the opportunity to contribute to a private version of *signifinder* which is still under development. To manage the multiple versions of the package, I used Git (Version 2.25.1) to access the latest package version and to maintain updated the package status. Git is a free and open-source version control system used to handle code projects efficiently: is used to track changes in the source code, enabling multiple developers to work together. The access to the main *signifinder* repository from which I cloned locally the *signifinder* was managed through a Gitlab interface.

Despite the data complexity and the thousands of computational operations required by the entire procedure, thanks to the efficiency of the data storage and the chosen tools the entire procedure can be run on a workstation. The one used for the development is equipped with Ubuntu 20.04.5 LTS type 64-bit, has 15,5 GiB of RAM and 8 processors Intel® Core™ i7-6700 CPU.

4 Results and Discussion

4.1 The analysis procedure

In the following paragraphs, the analysis procedure I applied and developed for the gene expression signature applications into the spatial transcriptomic experiments are presented step by step. The spatial transcriptomic analysis field is rapidly evolving and standard procedures and many analyses are not stable, therefore here below I also reported some points of discussion whenever different strategies can be considered.

4.1.1 The choice of the programming language of the analysis: the R language and the Bioconductor platform

In order to analyze spatial data through computational analysis and apply designed spatial packages it's necessary to store the data in a programmatic data object that encloses all the useful information compatible with the analysis strategy. To this purpose we used the R language jointly with packages available at the Bioconductor platform.
R is a free statistical software which provides a wide variety of statistical and graphical techniques. The R language is one of the most widely-used and powerful programming languages in bioinformatics, especially in the analysis of gene expression data.
Bioconductor is a free, open source and open development software repository that hosts a wide selection of analysis tools developed in the R programming environment. The project of Bioconductor is to develop and disseminate free open source software that facilitates rigorous and reproducible analysis of data from current and emerging biological assays. Bioconductor's main goal is to create a durable and flexible software environment by which statistical researchers can explore and interact fruitfully with shared data resources and algorithms. Bioconductor pipelines and analysis packages stand some rules and standardizations which aim is the simplification of data acquisition, management, transformation, modeling, combining different data sources and developing new modeling strategies. Being compatible with Bioconductor standards is generally considered a goal for the usability, reproducibility and distribution of a new software or an analysis procedure.

In the spatial transcriptomic context, Bioconductor proposes a data structure to store in one single object the necessary information of a spatial transcriptomic experiment coming from whatever technology. The constructed object is then used as a starting point to perform the vast majority of the analysis on spatial

transcriptomics data. The Bioconductor object dedicated to spatial transcriptomic experiments is called *SpatialExperiment*. *SpatialExperiment* is an extension of the SingleCellExperiment object, which is dedicated to store the single cell data.

4.1.2 Data loading: the Spatial objects data

After read alignment and quantification procedures, the reads, therefore the expression levels, can be assigned to genes, and following the spatial barcodes reads, and therefore the expression levels, can be assigned to each single spot with a specific spatial coordinate. The obtained expression matrix can be loaded into a *SpatialExperiment* object.

In a *SpatialExperiment* object are stored the spatial gene expression matrix, along with the row data and the column data, the spatial coordinates of the expressions, the image of the tissue slice and other image metadata.

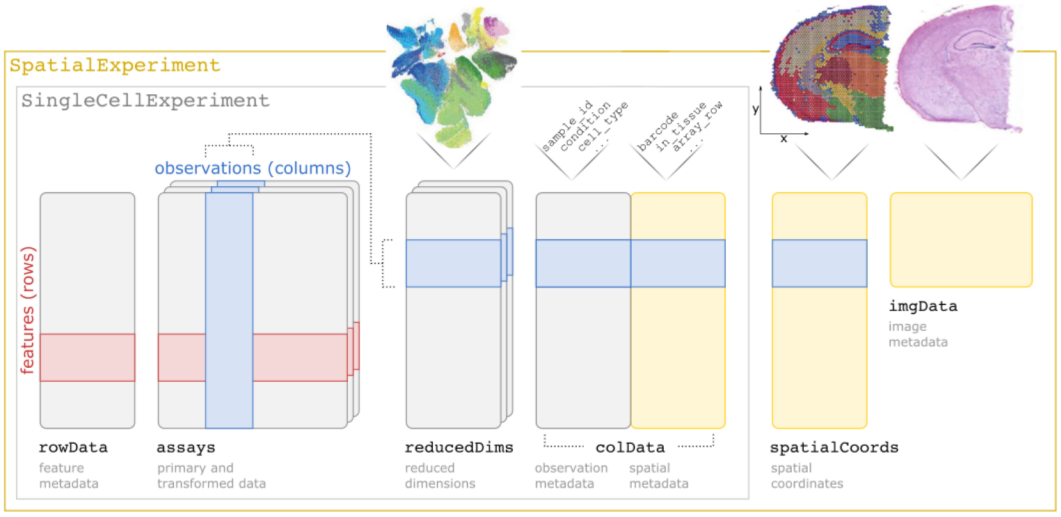


Figure 20: Bioconductor *SpatialExperiment* object structure. Figure credits to Righelli, SpatialExperiment: infrastructure for spatially-resolved transcriptomics data in R using Bioconductor, *Bioinformatics*, (2022) [27].

A *SpatialExperiment* can be manually constructed, loading manually every single piece of information in the specific data slot or, only for the 10x Genomics Visum data, can be used a specialized construction function.
The core of the object is the assay, or multiple assays, which consist in a matrix with gene names on the rows and spot identifiers on the columns fulfilled by the gene expression counts. As previously mentioned, spatial gene expression counts, as well as single cell expression matrix, have a high degree of sparsity, therefore the count values are saved in the dgcMatrix. The dgcMatrix is a class of sparse numeric matrices that allows to compress efficiently the aboundat

fraction of zeros count information. Then in the *SpatialExperiment* object there is a section for eventual row and column metadata, respectively *rowData* and *colData*. Additionally, linked to the *colData* there is also the *spatialCoords* section in which the spatial location of each spot in the image tissue is stored. Lastly, images and image metadata can be saved into the *SpatialExperiment* object.
In my pipeline, I decided to work with Bioconductor analysis and therefore to start the analysis with the loading of spatial transcriptomic data into the *SpatialExperiment* object.

As a worth of mention an alternative way of working with spatial experiments outside the Bioconductor environment is using the Seurat objects. Seurat is an R package designed for single-cell analysis and then adapted for the analysis and visualization of spatial transcriptomics data. Also Seurat proposes its own Spatial Seurat object class that can be automatically constructed for 10x Visium by a specific building function. As *SpatialExperiment*, the Seurat spatial object contains the expression matrix in the dgcMatrix format and different sections for all the information provided by the 10x Genomics.

4.1.3 Data preprocessing and normalization procedure of spatial transcriptomic data

The preprocessing step includes a spot selection procedure to filter out those spots with no or few genes quantified. Generally, the spots not covered by tissue do not show having reads; these include spots outside the tissue margins, holes or scratches of the tissue slice. Therefore, overlaying the spot coordinates of the expression matrix with the tissue image, all the spots not covered by tissue were removed from the analysis even if they show some reads because they are considered due to aspecific hybridization event.
As it happens for spots lying outside the tissue, it can also happen that spots covered by tissue show low read counts. This can due to (i) technical problem of the hybridization of RNA into the library probes of the spatial transcriptomic slide, or (ii) it can due to specific tissue/cell type or conditions which have few amount of mRNAs to be captured (e.g. necrotic or some connective tissues); these spots are respectively considered low quality data or unuseful outliers of the distribution and are removed. Empirical threshold based on data distribution and histologic information is generally chosen to identify these spots.

Once having the matrix with only the reliable spots, the normalization of read counts was performed. The techniques for the quantification of mRNA abundance introduce systematic sources of variation that can alter the amount of

mRNAs detected. Consequently, an essential first step in most mRNA-expression analyses is normalization, through which systematic variations are adjusted to make expression counts comparable across genes and samples or cells. There are two classes of normalization. One is the within-sample normalization to adjust gene-specific features, such as GC content and gene length. The other one is the between-sample normalization methods adjusted for sample-specific features, such as sequencing depth. The application of a between-sample normalization in the case of spatial transcriptomic data aims to remove artificial differences between total counts of spots, based on the fact that each spot has the same sequencing depth.

Compared to other transcriptomic data, the spatial data, as presented in the introduction of this thesis, present some peculiarities that can affect the success of the normalization procedure. The vast majority of spatial transcriptomics technologies, or at least the widely used such as the 10x Genomics Visium technology, is based on capture area of a fixed size divided equally in thousands of spots. Each spot retains the mRNA content of the overlaid tissue and the amount of cells within a spot. Thus, the transcriptional output given by a spot depends on the cell type, state, and the overall local tissue morphology. Specifically, the spot area of a 10x Visium experiment can contain from 5 to 20 cells; they therefore can show large differences in mRNA quantities. Given this precondition, which is an intrinsic and unsolvable issue of the spatial technology, normalization procedure in spatial transcriptomic data is still considered a challenge and even its application is a matter of debate. Given the similarity between the spatial and single-cell data distribution, standard procedures applied to spatial transcriptomic data the normalizations developed for single-cell data while others, such as Saiselet et al. in the work "*Transcriptional output, cell-type density, and normalization in spatial transcriptomics*" (*Journal of Molecular Cell Biology*, 2020) [30] explored and questioned if normalization in spatial data is necessary.

In fact, the authors stated that normalization could bias the differences in read counts between spots because the total counts per spot are biologically informative and reflect relevant quantitative and qualitative features of tissue morphology.

In our case study, as well as in all spatial and not-spatial transcriptomic data, different normalization procedures including the use of unnormalized data, can affect the results and this point will certainly require an in-depth analysis in the future.

4.1.4 The gene expression signatures: the *signifinder* R package

Signifinder is an under development R package that allows the user to compute a collection of cancer gene expression signatures on transcriptomics data. *Signifinder* can be found publicly available at https://github.com/CaluraLab/signifinder.

Currently, *signifinder* contains 46 signatures collected from tumor literature, providing an easy and fast way to obtain a fast implementation of signature scores for each sample. The standard analysis procedure of *signifinder* requires the submission of a gene expression data matrix from microarray or RNA-sequencing to signature dedicated functions. Through these functions signature scores will be calculated based on the expression of the genes and the algorithm defined by the signature. The output obtained is generally a single and/or multiple scores per signature that is added to the colData slot of a SummarizedExperiment object, the Bioconductor object used to store expression data. In *signifinder* additional functions are implemented to interpret, visualize and also to compare signatures scores between different samples.

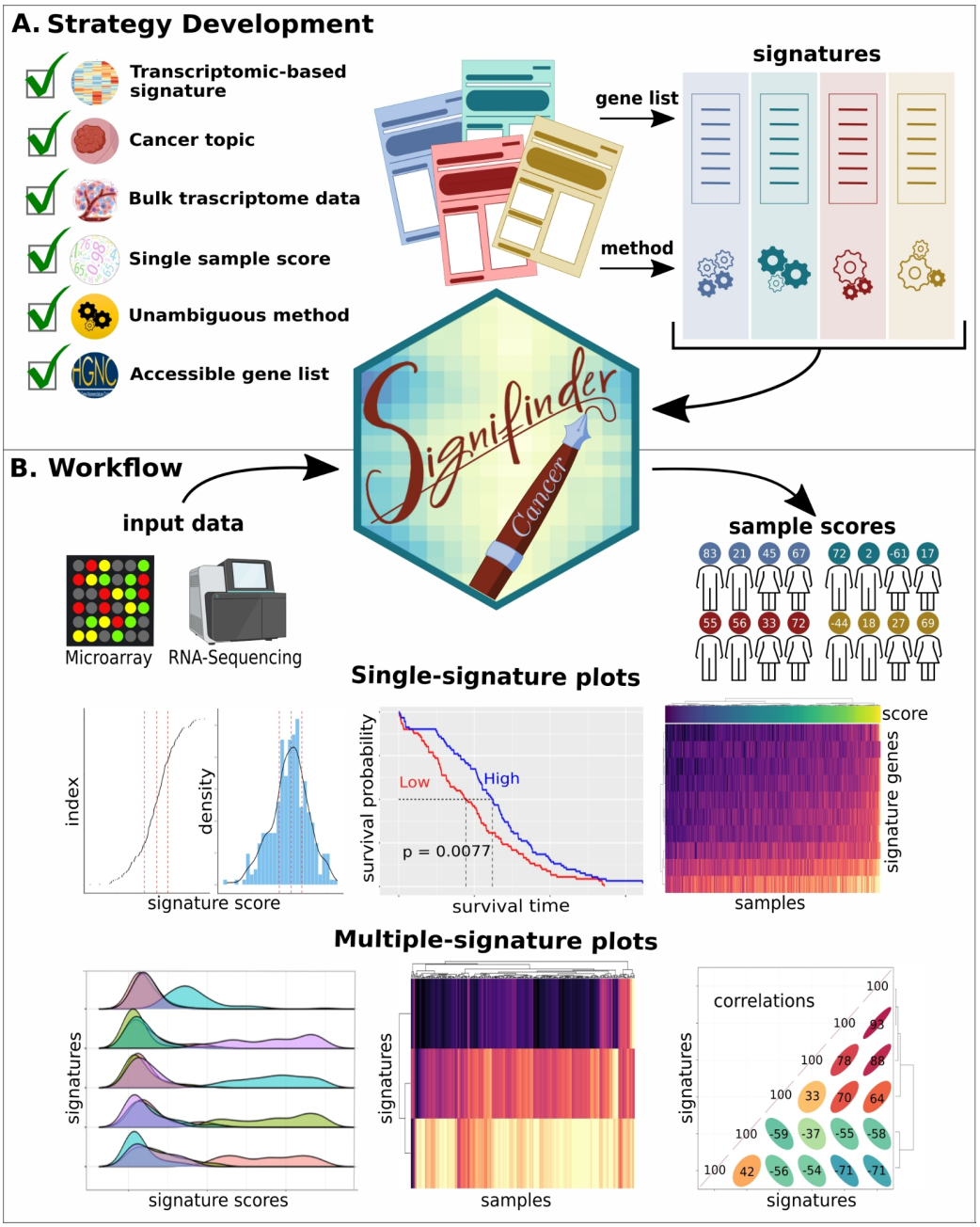


Figure 21: *signifinder* development and workflow overview. A: shows the schema for the *signifinder* development: following a set of stringent criteria, the lists of genes and algorithms for signature computation have been collected and implemented in dedicated functions. B: shows a typical *signifinder* workflow that starts with expression data and, without any other type of prior knowledge, the user can get and compare the sample scores for single or many signatures.

Signifinder and its collected signatures, has been developed to work with bulk transcriptome data, but there is no technical limitation to work with other types of transcriptome data such as single cell or spatial data. *Signifinder* takes the

expression matrix data provided (genes on the rows and samples on the columns), applies the signature algorithm and returns the resulting score for each sample. Equivalently, a score can be computed for each single-cell or quasi-single cell transcriptome returning multiple scores, one of each transcriptome or spatial spot.
To apply gene expression signature to spatial experiment I used a modified version of *signifinder* that can work with *SpatialExperiment* objects considering every single spot as if it was an independent sample.

4.1.5 Data visualization

In order to visually appreciate the spatial transcriptomics results in an actual spatial context I used the R package *ggspavis* (Version 1.2.0). This very recent package released in 2022 provides useful and clear visualization functions of spatially resolved transcriptomics data stored in *SpatialExperiment* format. One of the main functions of the package "*plotSpots*" allows the visualization of the spatial distributed spots in different dimensions and colors depending on their annotation, categorization or relative assigned scores.

4.1.6 Attempts to deal with the "zeros" problem

The results obtained from the previous signatures seem to be biologically meaningful and in agreement with what it's known about cancer. However, all the signatures I applied to the spatial transcriptomic data signatures have been designed and validated on bulk transcriptome data which substantially differs from spatial data. As previously explained, spatial data are characterized by a large amount of zero expression values which define the sparsity of the data. Those zeros can represent a true zero expression of a gene or it can be a consequence of a miss detection of an expressed gene due to technical limitations. All together the zeros values represent 75% of the spatial data and the signatures, designed for bulk data, do not take into account this aspect in the computation of scores, which can bring mathematical challenges. For example, for those signatures that involve the computation of the geometric mean the amount of zeros expression values is highly relevant. The geometric mean of a n set of values is given by the nroot of the multiplication of those n numbers, therefore each zero value should be discarded in the set of expression values considered. As a consequence, this is not relevant if the dataset has a very low percentage of zero expression, otherwise, if the amount of zero is high a large amount of data should be discarded and the obtained score will be strongly biased. On the other hand, if the distribution of zeros is equivalent in all the

spots belonging to the same type of tissue, the resulting score might be comparable and relevant anyway. As mentioned in the background session, some approaches that aim to impute artificial zeros have been proposed for spatial data. Generally, they try to reduce the number of zeros and increase the counts values of each spot leveraging the spatial information, resulting in the implementation of spatial smoothing algorithm. As well, I tried to apply a similar adjustment on the count values of the spatial gene expression data under consideration. The adjustment applied is way simpler than the complex approach implemented by the spatial packages as *stLearn* or *spaGCN* but it can be an approach to handle spatial data.

Two simple ideas have been implemented to see if spatial data can actually gain improvements in terms of zero abundance and signature score distribution. The consecutive rows of capture spots that define the grid of Visium technology are staggered one to the other, arranged in a way that makes it easy to define triads of close spots as shown in the figures below.

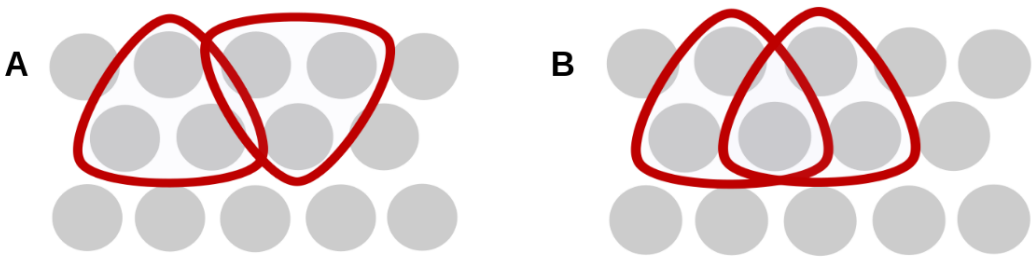


Figure 22: A. non overlapping selection of triad of spots. B. overlapping selection of triad of spots.

In the first adjustment type (Figure 22A) the expression profiles of the spots in a triad are summed up and the resulting profile is assigned to each of the spots constituting the triad. The following step is to consider the subsequent upside down triades of spots, repeat the procedure, and so on for the whole grid of spots.
The second adjustment type (Figure 22B) mimics the idea of the sliding window: the first step is considering one triad of spots, then the three expression profiles are summed up and assigned only to the spot on the top of the triangle. The next triad is defined by moving one spot further in the line and repeating the sum and the assignment along the whole grid.
An open question remains on the handling of the edge spots in the adjustment of the expression counts. The methods applied take for granted that one spot has its neighbor triad with transcriptional information, but in the case of edge spots along the border of the tissue this is not a valid approach. It may seem a marginal

problem but the number of edge spots rapidly increases when, like in the case presented below, there are empty or excluded areas inside the tissue. The challenge occurs when there are no close profile spots to sum and the edge spots remain unchanged by the adjustment with a lower total count and highest number of zeros.

4.2 The case study

4.2.1 Human ductal breast cancer

Breast cancer is a disease in which breast cells starts to hyperproliferate. A healthy female breast is made up of 12-20 sections called lobes, each one made up of many smaller lobules which define the glands that produce milk in nursing women. Lobes and lobules are connected by milk ducts that carry the milk to the nipple. These breast structures are generally where the cancer begins to form. Less commonly, breast cancer can begin in the stromal tissues, which include the fatty and fibrous connective tissues of the breast.
Breast cancer can arise from the cells of the epithelium of the ducts or from the cell of the lobules in the granular tissue of the breast. Initially, the cancerous growth is confined to the duct or lobule, defined as "in situ", having the potential to spread. Over time, the in situ cancer may progress to invade the surrounding breast tissue becoming an invasive breast cancer. It can spread in the nearby lymph nodes originating regional metastasis or to other organs of the body establishing distant metastasis. There are two common types of breast cancer: invasive ductal carcinoma and invasive lobular carcinoma depending which types of cells start to become neoplastic.

Breast cancer is one of the most common cancers worldwide and, although it can affect both women and men, only 0.5 -1% of breast cancer diagnoses occur in men. Breast cancer is mainly diagnosed in women over 40 years and its risk increases with age, obesity and unhealthy life habits as harmful use of alcohol and tobacco. Invasive ductal carcinoma is the most common type of breast cancer, making up nearly 70- 80% of all breast cancer diagnoses both in women and men. Most breast cancers are sporadic, but 5% to 10% of the cases are thought to be hereditary caused by a predisposition principally due to mutations in the tumor suppressor genes BRCA1 and BRCA2.

4.2.2 Histologic reading of the sample

10x Genomics data includes a high definition histological image of the tissue slice used in the spatial analysis, in this case a breast carcinoma, reported below.

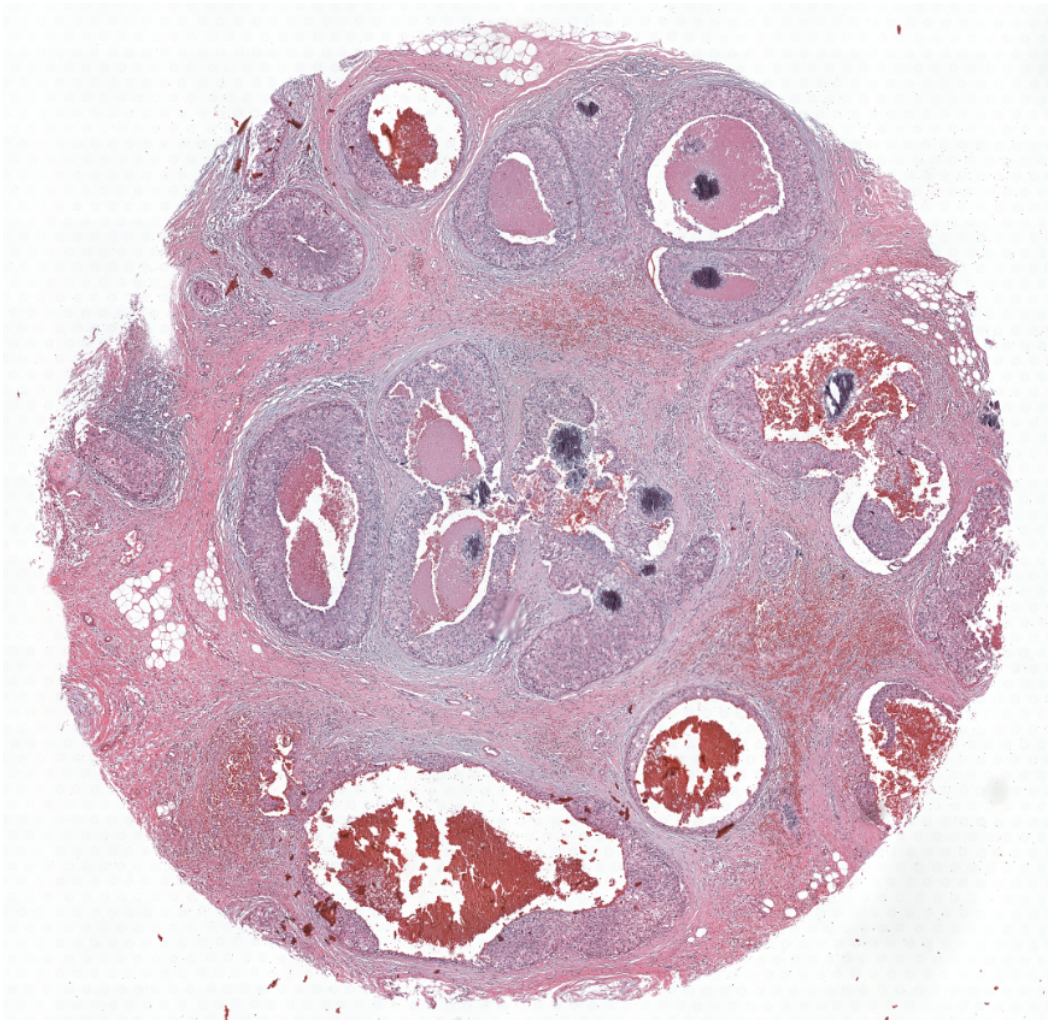


Figure 23: H&E image provided by 10x Genomics.

We submitted the image for a professional histopathological reading to the anatomopathologist Gennaro Esposito (*IOV*, *Istituto Oncologico Veneto* I.R.C.C.S., Padova). Based exclusively on the morphological characteristic of the cells, tissue and cells can be categorized. The healthy breast is mainly composed of stromal cells which are fibroblasts, immune cells, adipocytes and endothelial cells, along with extracellular matrix (ECM) components, the most abundant of which is collagen I.

In Figure 24 we can find the areas with specific tumor properties, mainly delimiting the tumoral tissue from the healthy regions, they are highlighted in red, blue and black lines.

(i) The multiple neoplastic areas, highlighted in red**,** can be seen in the sample. In this section the tumor areas are localized inside the duct, as expected by this type of tumor since its origin in the basement membrane of a breast duct and it generally invades the inner part of the duct.

(ii) Tumor masses are surrounded by fibrous tissue, delimited in blue, which is identified as cancer associated fibroblasts (CAFs). CAFs are generally found in the cancer stroma and are the major component of it. Fibroblasts around tumor mass contribute to its proliferation, invasion capabilities and metastasis through the secretion of several growth factors, cytokines, chemokines, and degradation of extracellular matrix (ECM) proteins. In the rest of the tissue have been recognised adipocytes, blood vessels and some small aggregation of lymphocytes.

(iii) As can be appreciated in the figure, multiple necrotic areas are present and are indicated in black. In the figure the necrotic cells are mainly located inside the regions labeled as tumor, which is a peculiar feature of advanced solid tumors and is associated with poor prognosis of cancer patients. Generally the inner cells of solid tumors display insufficient blood irroration which is translated into oxygen and nutrient deprivation leading to necrotic death. Recently necrosis has been recognised not as an accident consequence of tumor growth, but as a programmed cell death with a tumor-promoting potential. Other tissue structures have been recognised as adipocytes, lymphocytes and blood vessels respectively indicated in Figure 24 in green, light blue and orange.

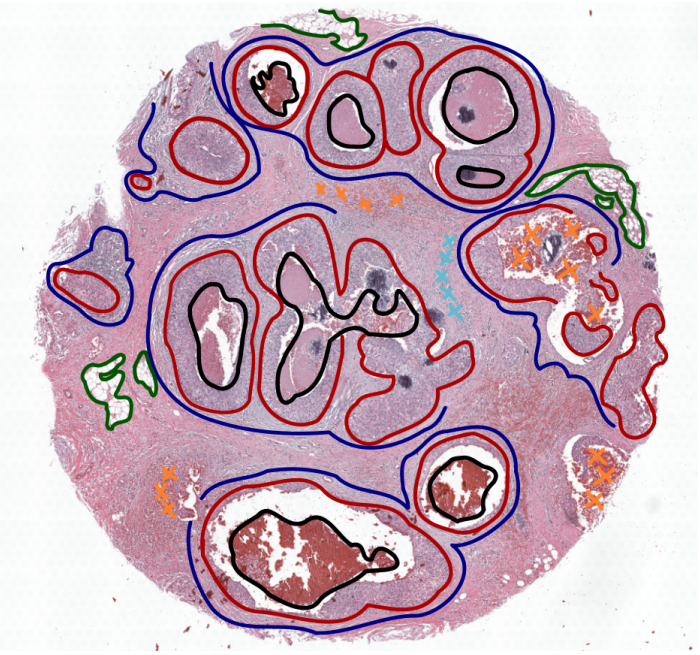


Figure 24: Professional histopathological reading of the H&E image, with colors to indicate different areas of the biopsy: neoplastic areas (red), necrotic areas (black), cancer associated fibroblast (blue), adipocytes (green), lymphocytes (light blue), blood vessels (orange).

Using CellRanger, the proprietary software of the 10x Genomics, we can easily translate the histologic annotations to each spot of the spatial transcriptomic grid, along with the tissue border. The result of this procedure can be found in Figure 25. Thus, each spot of the capture area has been annotated as: tumor, necrosis, lymphocytes, CAFs, blood vessels, adipocytes and stroma. The assignment has been performed manually for each spot, and for those spots located on the edges of different adjacent sections the label was defined by the

structure that covers the major area of the spot. The annotation "stroma", which enclose different cell types, was used when multiple cell types underlie the same spot.
This procedure gives an "identity" to each spot of the capture area, in this way this information can be automatically read and considered by the analysis procedure.

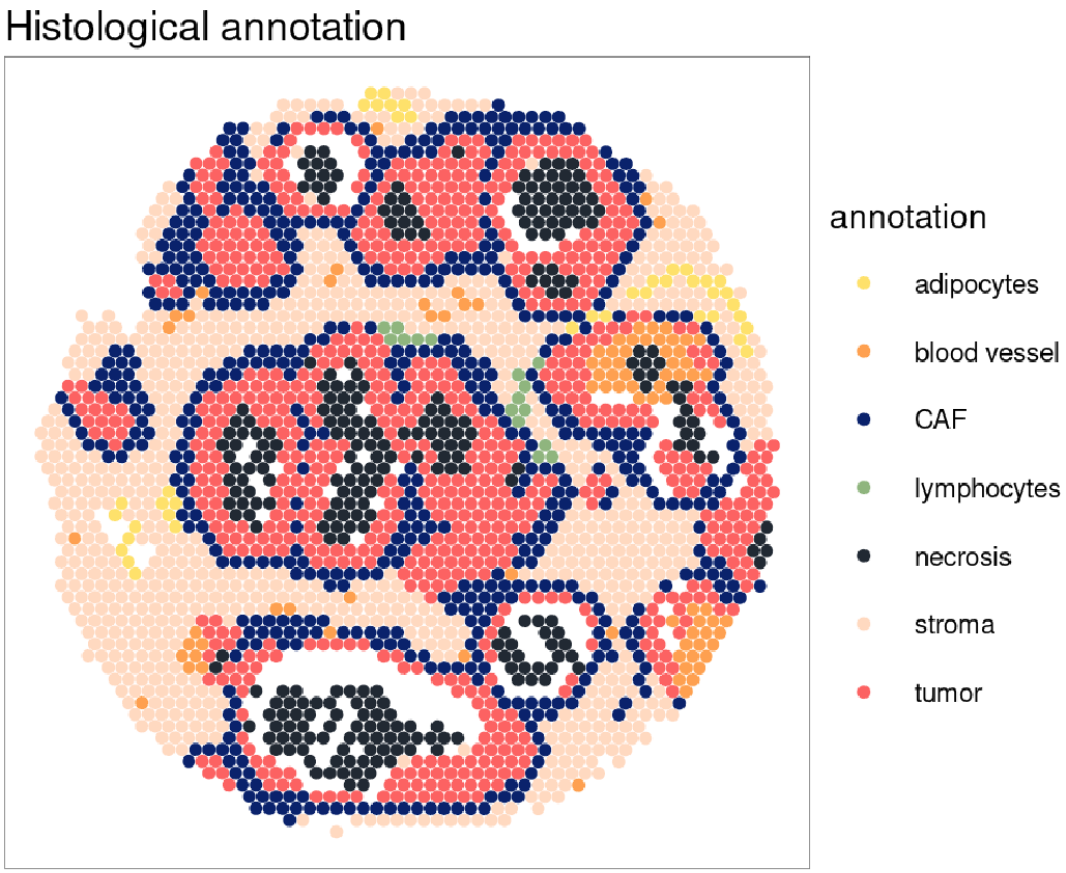


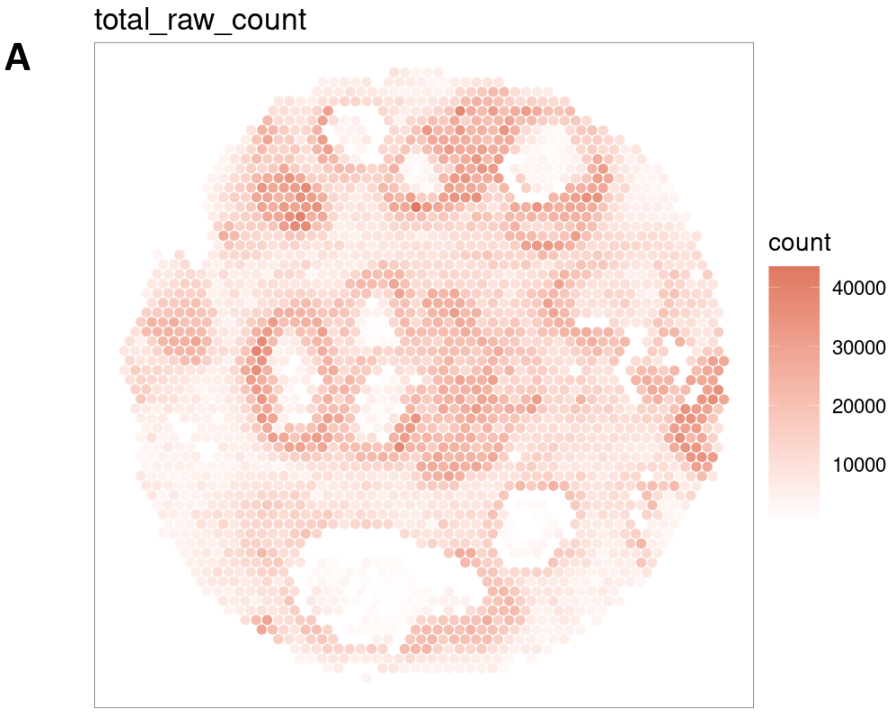
Figure 25: Manual transfer of the histological annotation of the tissue slice of the 10x Ductal Breast Cancer Sample into the spots of the capture area using the 10x software Loupe Browser.

4.2.3 Expression data analysis

The spots not covered by tissue do not generally show RNA counts, they appear as lacking in the figures and were not considered in the analysis, these include spots outside the tissue margins, and holes or scratches of the tissue slice. Sometimes, these spots show very low tissue counts due to aspecific hybridization of RNA into the library probes; these spots are generally considered of low quality and removed. Specifically in this sample we have 2518 spots showing at least one count in one gene. The vast majority of these low count

spots correspond to necrotic areas, dead cells do not have mRNAs or the vast majority is too degraded to be linked by the slide library probes. Thus all the spots with very low counts or annotated as necrotic, have not been taken into account in the analysis. The small amount of transcript sampled in that area can also be due to the possible diffusion of mRNAs from other cells during the permeabilization of the tissue. As a consequence, those spots have been removed before normalization of the row counts.

The normalization applied is the one suggested for single cells transcriptome data that is applicable to the SpatialExperiment object, which consists in the log normalization of the row counts proposed by the single-cell analysis package *scater*. Normalization aims to remove the differences between total counts of spots imputed to the technical variation in RNA capture, thus obtaining comparable spots in terms of counts and gene expression levels.



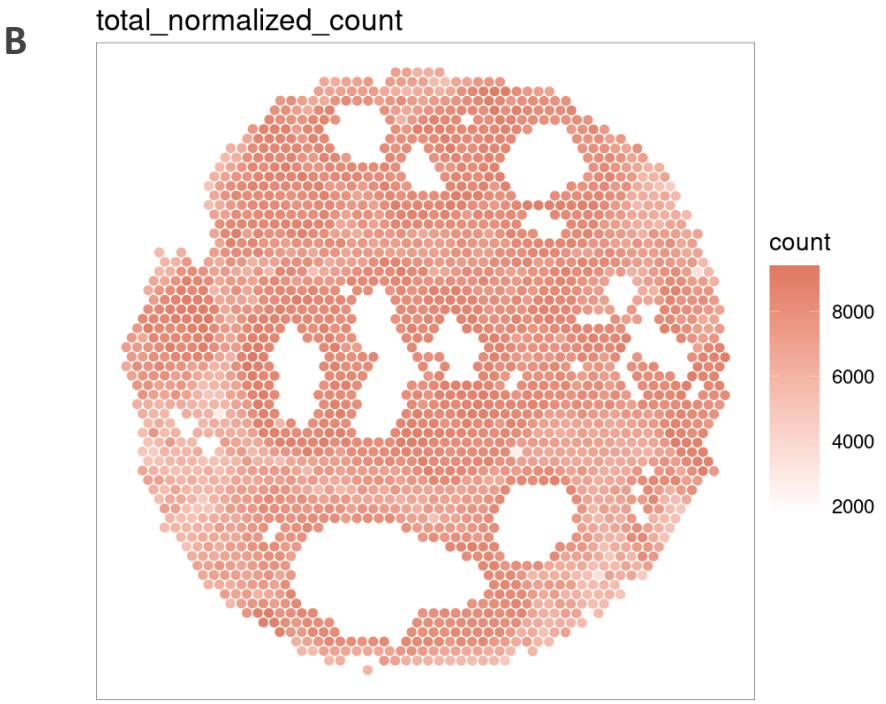


Figure 26: A. total amount of gene expression counts per spot before normalization. B. total amount of gene expression counts per spot after normalization.

It is clear in the plots above (Figure 26) how the normalization applied makes the total count of each spot more uniform across the tissue. However the total counts can not be forced to be identically distributed between all the spots because the density of cells in the tissue is not homogeneous, while by construction the spot areas always have the same size, therefore spots covering different numbers of cells could collect largely different amounts of transcripts.

4.2.4 Computation of pancancer signatures

Signifinder contains a collection of tumor-specific and pancancer signatures (signatures that can be used in all types of cancers). All the signatures that I applied to this Visium experiment of breast cancer are listed and described in table 1, they are all the pancancer signatures.

Signature	Topic	Description	Reference
CellCycle_Lundberg	cell cycle	It is a representative of general cell-cycle activity and could be applied to any tissue sample. Higher scores represent a worse prognosis.	[31]

Signature	Topic	Description	Reference
MitoticIndex_Yang	cell cycle	The mitotic-index is constructed from genes that have been highly validated as being cell proliferation markers. The score reflects the fraction of dividing cells in a sample and can be used as a predictors of normal/cancer status.	[32]
CIN_Carter	chromosomal instability	The score characterizes aneuploidy in tumor samples based on coordinated aberrations in expression of genes localized to each chromosomal region. Higher the score higher the total level of chromosomal aberration. Net overexpression of this signature was predictive of poor clinical outcome in six cancer types.	[33]
Hypoxia_Buffa	hypoxia	A highly prognostic signature. The score increasement reflects hypoxia activity.	[34]
ImmunoScore_Roh	immune system	The score is based on expression of genes involved in cytolytic markers, HLA molecules , IFN-γ pathway genes, chemokines, and adhesion molecules. It is used to investigate immune activation in tumor microenvironment, higher the score higher the immune system activation in relation to tumor rejection.	[35]

Table 1: Computed signatures.

Once the score signature function is applied on spatial data matrix we obtain a score for each expression profile of each spot, which is stored in the colData of the *SpatialExperiment* object.

Those scores can be plotted on the "spot" image codified by a color scale in order to obtain a visual representation of the distribution of the results all over the tissue slice.

In the following paragraphs a selection of signatures mapped on the spatial data, along with the signature explanation and results discussion.

Cell cycle signature

The main distinctive character of tumor tissue is the uncontrolled growth of cells, so the dysregulation of the cell cycle is one of the key markers of the tumoral tissue. Lundberg et al. [31] proposed a cell cycle signature based on 463 cell-cycle related genes collected from KEGG, HGNC and Cyclebase. The cell cycle signature score (CCS score) is calculated considering the mean expression values of the selected genes, and is used in pan-cancer analysis and represents the marker of cell cycle activation and cell proliferation.

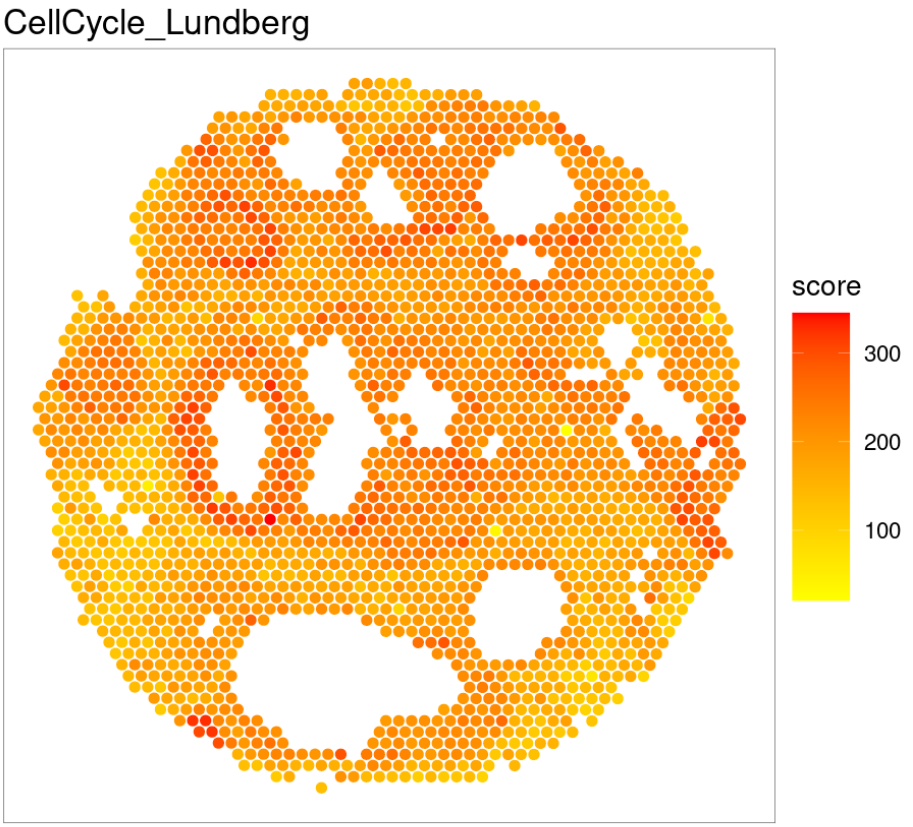


Figure 27: CellCycle signature score distribution all over the spots.

The distribution of the CellCycle score all over the tissue spots depict multiple intense-red regions and other more extended orange-yellow areas. As we expected, the higher scores are attributed to those spots covering the tumoral area, while the rest of the stroma displays an overall lover resulting scores highlighting an increased cell cycle activity in the fast replicating tumor cells.

Mitotic index signature

Cancer risk in somatic tissue is correlated to the rate of stem cell division, thus a marker able to approximate stem cell divisions in a tissue can be useful to individuate high probability of cancer triggering. Yang et al. [32] developed a mathematical model able to approximate mitotic clock in both normal and cancer tissue, demonstrating that this mitotic-like clock is universally accelerated in cancer, precancerous lesions and normal epithelial cells exposed to a major carcinogen. The mRNA based mitotic signature is calculated on the mean of the expression level of 9 genes, related to the proliferation, and reflects the rate of division in the tissue.

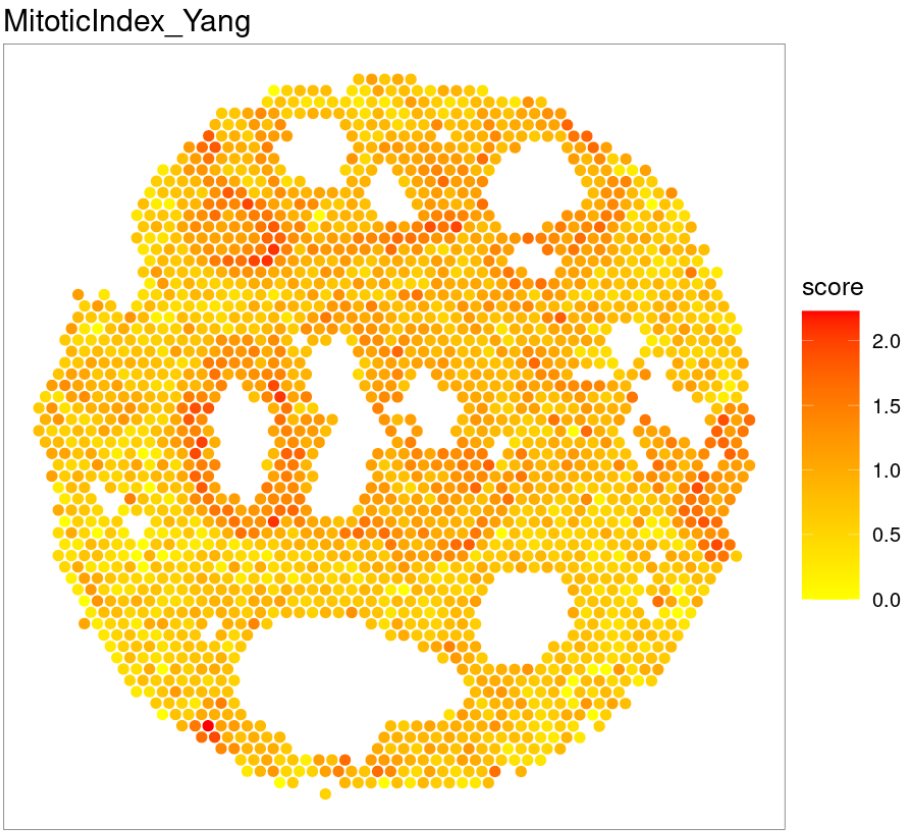


Figure 28: Mitotic index signature score distribution all over the spots.

The distinctive characteristic of tumors is the abnormal proliferation of cells which, as a direct consequence, increases the number of accumulated cell divisions. Therefore, in the resulting plot of the *mitotic index* above, on average, the spots with higher scores are those labeled as tumor.

Chromosome instability signature

One of the most consistent characteristics of human solid tumors is chromosomal instability (CIN) which results from errors in chromosome segregation during mitosis. Carter et al. [33] constructed the CIN signature based on the expression level of genes consistently associated with aneuploidy. The authors reasoned that aneuploidy is a consequence of chromosomal instability and so aneuploidy correlated genes might provide insight into molecular mechanisms underlying chromosomal instability. Genes found correlated to aneuploidy by the authors have been ranked and the top 70 correlated genes are used in the calculation of the CIN signature score. This signature is based on the normalized sum of the gene expression count of those 70 genes, mainly involved in the faithful replication and segregation of chromosomes.

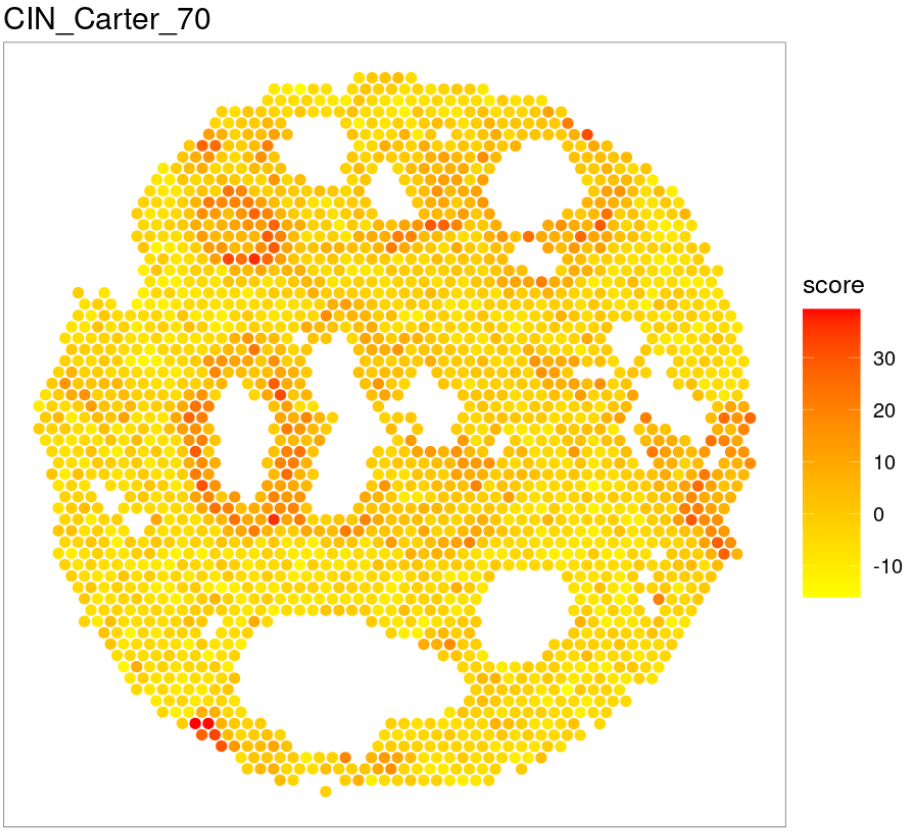


Figure 29: Chromosomal instability signature score distribution all over the spots.

As the mitotic index, also the chromosomal instability is one direct consequence of the uncontrolled proliferation of tumor cells, indeed the above plot displays a similar score distribution with the mitotic index plot. In Figure 29 can be appreciated multiple dark red spots distributed in a subpart of the tumor

annotated area highlighting a non homogenous distribution of this signature inside the tumor mass and tantalizingly suggesting that tumor cells can be divided in two classes of stable and unstable tumor cells.

Hypoxia signature

Low level of oxygen is another major feature of solid tumors that favors tumor progression. Hypoxia, both in normal and neoplastic tissue, induces a molecular response that stimulates the growth of new vasculature, essential for nutrient supply and dissemination of neoplastic cells. Buffa et al. [34] propose a Hypoxia signature built on 49 genes that the authors found consistently co-expressed with the low level of oxygen in multiple cancers. The signature score is based on the median expression of those selected genes and is used to investigate the hypoxia level in the tissue.

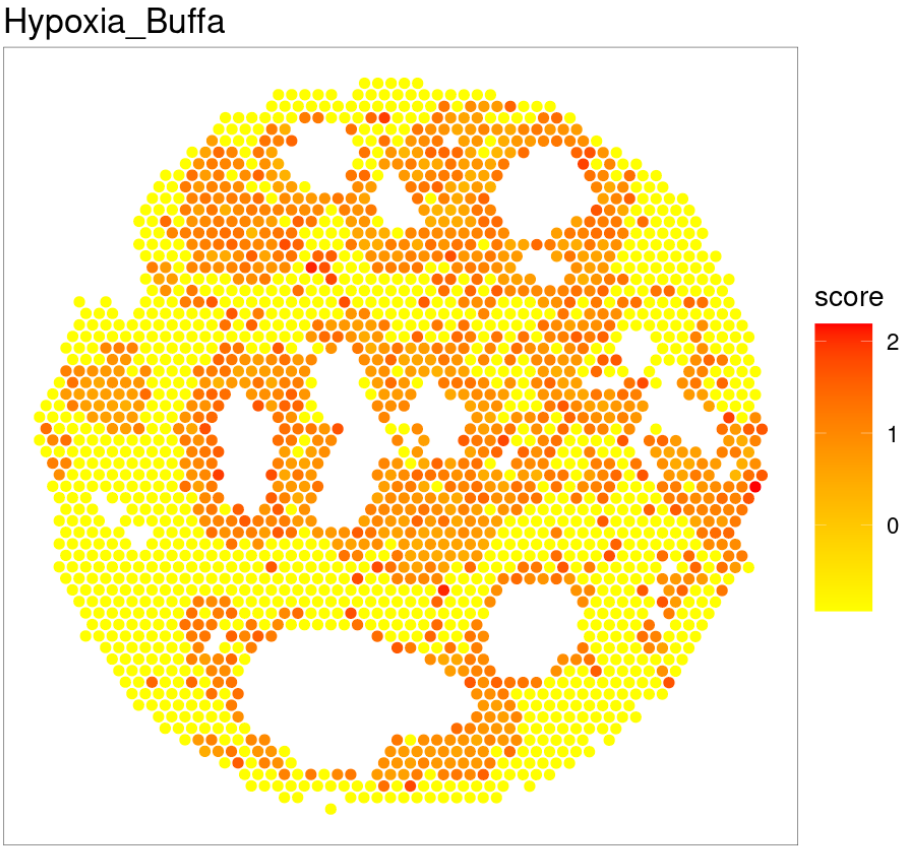


Figure 30: Hypoxia signature score distribution all over the spots.

The hypoxia signature score shows a clear separation between tumor and non-tumoral tissue; indeed the results show almost the total of the tumors sports with high scores. Tumor tissue of this sample seem to be strongly

characterized by low levels of oxygen and this is also confirmed from an histologic point of view by the large areas of tumor necrosis. Low level of oxygen could stimulate angiogenesis and the block of which is the target of multiple anticancer adjuvant therapies. Hypoxia sign is calculated on the median expression of 9 hypoxia related genes which, supposedly, are activated only in the tumor tissue and the CAFs region surrounding it.

Immune system signature

Immune system and cancer present a complex relationship that defines the balance between the immune surveillance and immune escape, which leads to cancer progression. The Immune signature proposed by Roh et al. [35] is based on 41 immune related genes selected for melanoma. This score depends on the expression of genes associated with immune activation in the TME and can be used to evaluate the activation state of the immune system in cancer. The immune signature, valid as pan-cancer signature, is defined by the geometric mean of gene expression counts of those 41 immune related selected genes.

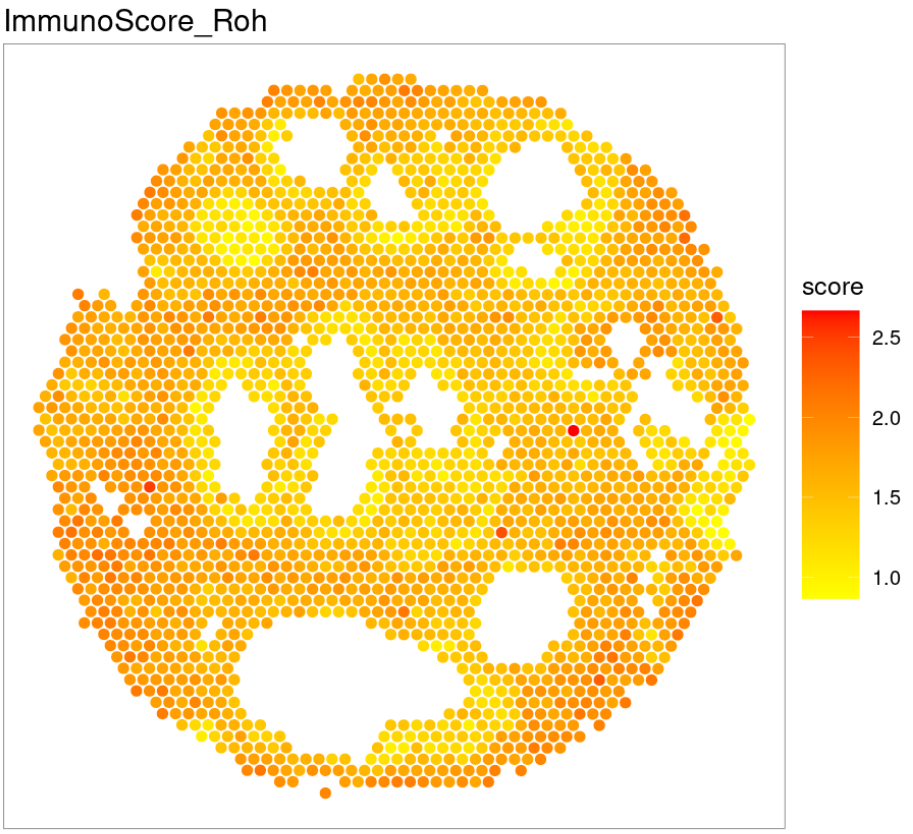


Figure 31: ImmunoScore signature distribution all over the spots.

In the figure above is reported the distribution of the immunoScore in which the lower score region corresponds to the neoplastic area, while the rest of the spots shows a higher score level. The ImmunoScore is calculated on the expression of immune related genes which reflects the immune response of the tissue to the tumor offense. Generally, a higher and active immune response, especially within the tumor tissue, is linked to a good prognosis. On the other hand, in this case the tumor area is characterized by a lower ImmunoScore than the surrounding tissue, which can be a hint of an advanced tumor state, in line with all the other observations. The study of localization and activity of the immune system cells are of paramount importance especially in the context of immunotherapies, having the information that immune cells are present and active outside of the tumor borders can be an interesting information that can be used in a clinical path for the immunotherapy choice.

4.2.5 Smoothing approaches on the case study data

As mentioned before, spatial gene expression data are fulfilled by zero counts. The large amount of zero counts, true or artificial, covers 75% of this case study data. The total quantity of zeros define the high level of sparsity of the data which could bring challenges in the application of already existing analysis tools not specifically designed for spatial data. The idea to exploit spatial information to overcome the zero amount problem has been already implemented through different and complex approaches integrating both morphological information and spatial location.
To my case study I implemented and applied two different smoothing approaches, both based on the selection of adjacent tris of spot and the union of their total counts, as represented in the preceding Figure 22. The first one entails the selection of a triangle of three spots, the sum up of their counts and then the reassignment of the resulting merged counts to all three sports. This method is defined as the non-overlapping method because one spot belongs to only one triad. On the other hand, in the second adjustment method, indicated as the overlapping triad adjustment, one spot belongs to multiple triads, mimicking the idea of the sliding window.

Once applied both two smoothing methods the calculation of the signatures have been performed again and below it's reported the chromosomal instability signature score distribution.

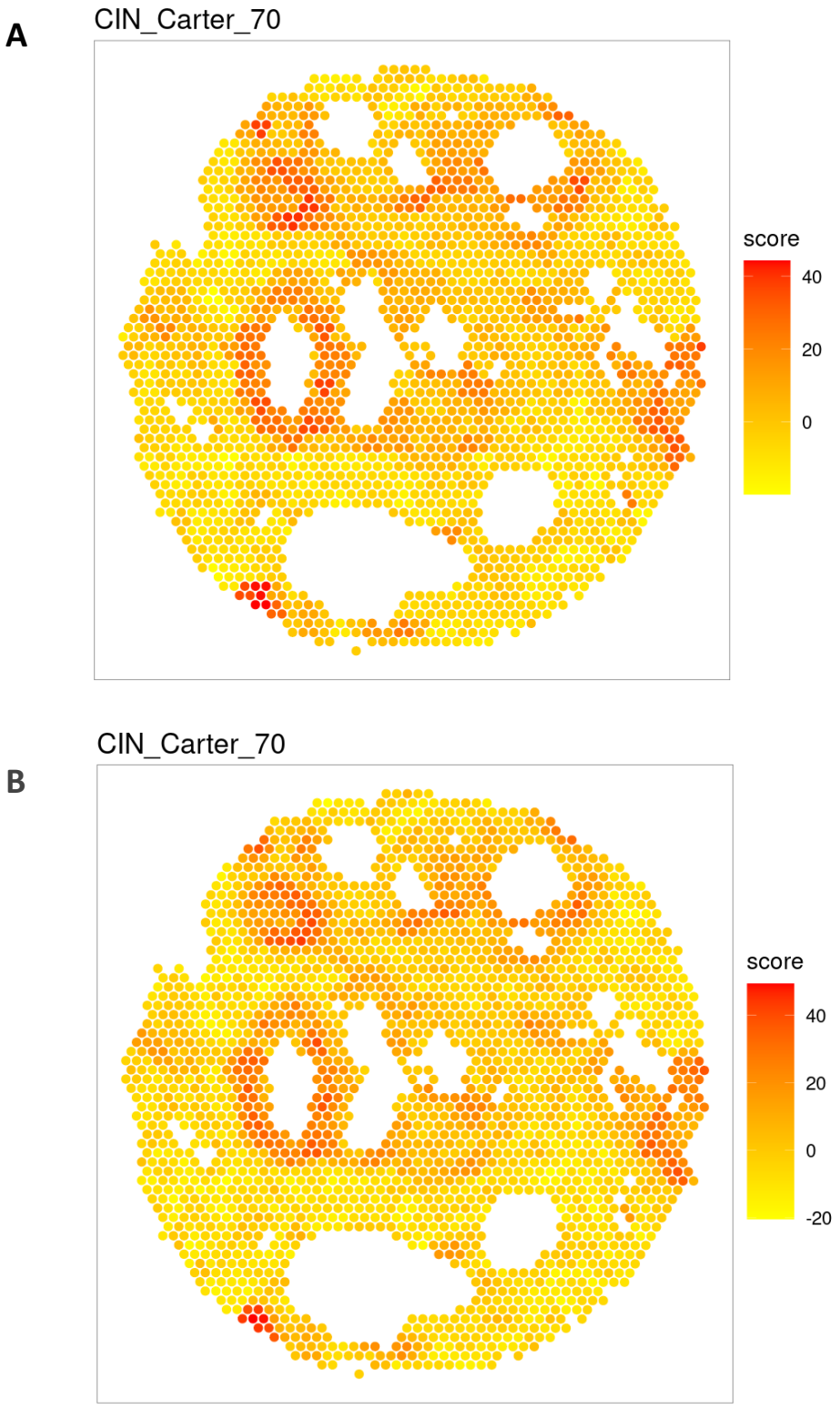


Figure 32 : A. Chromosomal instability signature score distribution after the non overlapping triad adjustment counts method. B. Chromosomal instability signature score distribution after the overlapping triad adjustment counts method.

With both two methods applied the overall distribution of the signature score remains unchanged which is a sign that the changes applied have not distorted the data. Then, in both plots it can be appreciated that the area with higher score seems to be more extended compared with the unsmoothed data, which is a common result obtained from two different "adjustment actions".
In the first case we select three spots and we merge their information without sharing it with the other neighbors spots, so basically we are incorporating a triad of spots, increasing the total spot count but decreasing the resolution of the transcriptional pictures. Indeed, we obtain a lower resolution averaged score distribution, in which equal score tirad spot can be easily recognised.
In the second case we sum up three close spot transcription profiles allowing the sharing of counts between subsequent triads. The reassignment of the merged profiles is performed to only one spot at the time, aiming to maintain the initial resolution of the technology. The direct result of this sharing is a more blended appearance of the score distribution that makes the high score areas look more extended.

In terms of zero abundance reduction both methods are almost equivalent: on average 12950 genes have zero counts values per spot before adjustment, instead after it the average zero count genes per spot decrease to 9950. Differently from what we expect, the reduction in zero abundance is not that strong because, generally, a gene with zero count in a spot tends to have zero count also in its neighbors spots.
Although those are preliminary results, these approaches are helpful in the reduction of zeros in the spatial gene expression data. Currently, we are working on the application of the smoothing approaches previously presented in the introduction chapter, that seems to strongly solve the problem of the zeros abundance.

5 Conclusions and Future Perspectives

The main topic of my thesis project is the new spatial technology approach for which several analysis tools are under development. The current diffusion of spatial transcriptomics technologies is bringing up to light all the possible advantages of its application, both in research and in clinical practice. For example, the ability to read tumor biopsies, not only based on the morphology of the tissue, but also based on the transcriptional information, could provide a better interpretation of tumor subtypes. Proper recognition of tumor subtypes leads to precise stratification of patients and adequate treatment selection. However, intuitive and user-friendly tools that can allow the medical personnel to perform this analysis are still missing.

This thesis project represents one first step into the application and adaptation of gene expression cancer signatures on spatial transcriptomics data, aiming to provide effective interpretation strategies through knowledge and algorithms designed for bulk analysis. To do this, we applied *signifinder*, an R package for the computation of a compendium of gene expression cancer signatures. The results obtained by the application of *signifinder* are biologically meaningful and display an overall agreement between each other. Altogether, the results of the signature highlight an advanced tumor state with poor prognosis markers, such as the high level of hypoxia and the low immune response in the tumor area.

Currently, *signifinder* provides several functions for the computation of signatures on bulk data with additional functions for the interpretation and comparison of final scores. The future step will be the full adaptation of *signifinder* to work with spatial data by the integration of analysis functions specifically designed to work with it. For example, the addition of smoothing functions in the package could give the user the possibility to adjust the count distribution, compare the impact of it and evaluate the best results. As previously presented, few articulated methods that integrate the morphology of the tissue slice in the smoothing method have been already proposed in the *stLearn* and *spaGCN* packages [21, 23]. One easier way to reproduce this type of approach could be the integration of the histological annotation of spots covering the tissue as a criterion to share the counts of a spot with its neighborhood.

Anyway, several open questions remain unsolved. First of all, about the normalization step there are differing opinions, arguing if it is actually required in the analysis of spatial data [30]. In addition, smoothing methods aim to overcome the issue of zero abundance and low counts values through the

adjustment of total counts, leveraging spatial information. Regardless how reasoned and motivated the adjustment may be, it will inevitably remain a modification of the original data with the possible introduction of artifacts. Linked to the adjustment strategy, another unsolved issue remains the handling of border spots which are not surrounded by other tissue spots.

All things considered, due to the promising advantages and cutting edge applications of spatial transcriptomics technologies, it is, with no doubt, worthy to persevere in the development of spatial analysis and interpretation softwares.

References

[1] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674, 2011.

[2] Joseph R Nevins and Anil Potti. Mining gene expression profiles: expression signatures as cancer phenotypes. *Nature Reviews Genetics*, 8(8):601–609, 2007.

[3] Jyothi Subramanian and Richard Simon. What should physicians look for in evaluating prognostic gene-expression signatures? *Nature reviews Clinical oncology*, 7(6):327–334, 2010.

[4] Maryann Kwa, Andreas Makris, and Francisco J Esteva. Clinical utility of gene-expression signatures in early stage breast cancer. *Nature reviews Clinical oncology*, 14(10):595–610, 2017.

[5] Marco Gerlinger, Andrew J Rowan, Stuart Horswell, James Larkin, David Endesfelder, Eva Gronroos, Pierre Martinez, Nicholas Matthews, Aengus Stewart, Patrick Tarpey, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl j Med*, 366:883–892, 2012.

[6] Andrew L Ji, Adam J Rubin, Kim Thrane, Sizun Jiang, David L Reynolds, Robin M Meyers, Margaret G Guo, Benson M George, Annelie Mollbrink, Joseph Bergenstrahle, et al. Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. *Cell*, 182(2):497–514, 2020.

[7] Tong Fu, Lei-Jie Dai, Song-Yang Wu, Yi Xiao, Ding Ma, Yi-Zhou Jiang, and Zhi-Ming Shao. Spatial architecture of the immune microenvironment orchestrates tumor immunity and therapeutic response. *Journal of hematology & oncology*, 14(1):1–25, 2021.

[8] Carlos Carmona-Fontaine, Maxime Deforet, Leila Akkari, Craig B Thompson, Johanna A Joyce, and Joao B Xavier. Metabolic origins of spatial organization in the tumor microenvironment. *Proceedings of the National Academy of Sciences*, 114(11):2934–2939, 2017.

[9] Anjali Rao, Dalia Barkley, Gustavo S Franca, and Itai Yanai. Exploring tissue architecture using spatial transcriptomics. *Nature*, 596(7871):211–220, 2021.

[10] Samuel G Rodriques, Robert R Stickels, Aleksandrina Goeva, Carly A Martin, Evan Murray, Charles R Vanderburg, Joshua Welch, Linlin M Chen, Fei Chen, and Evan Z Macosko. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434):1463–1467, 2019.

[11] Sanja Vickovic, Gokcen Eraslan, Fredrik Salmen, Johanna Klughammer, Linnea Stenbeck, Denis Schapiro, Tarmo Aijo, Richard Bonneau, Ludvig Bergenstrahle, Jose Fernandez Navarro, et al. High-definition spatial

transcriptomics for in situ tissue profiling. *Nature methods*, 16(10):987–990, 2019.

[12] Yang Liu, Mingyu Yang, Yanxiang Deng, Graham Su, Archibald Enninful, Cindy C Guo, Toma Tebaldi, Di Zhang, Dongjoo Kim, Zhiliang Bai, et al. High-spatial-resolution multi-omics sequencing via deterministic barcoding in tissue. *Cell*, 183(6):1665–1681, 2020.

[13] Xiaonan Fu, Li Sun, Jane Y Chen, Runze Dong, Yiing Lin, Richard D Palmiter, Shin Lin, and Liangcai Gu. Continuous polony gels for tissue mapping with high resolution and rna capture efficiency. *BioRxiv*, 2021.

[14] Chun-Seok Cho, Jingyue Xi, Yichen Si, Sung-Rye Park, Jer-En Hsu, Myungjin Kim, Goo Jun, Hyun Min Kang, and Jun Hee Lee. Microscopic examination of spatial transcriptome using seq-scope. *Cell*, 184(13):3559–3572, 2021.

[15] Ao Chen, Sha Liao, Mengnan Cheng, Kailong Ma, Liang Wu, Yiwei Lai, Jin Yang, Wenjiao Li, Jiangshan Xu, Shijie Hao, et al. Large field of view-spatially resolved transcriptomics at nanoscale resolution. *BioRxiv*, 2021.

[16] Helga Bergholtz, Jodi M Carter, Alessandra Cesano, Maggie Chon U Cheang, Sarah E Church, Prajan Divakar, Christopher A Fuhrman, Shom Goel, Jingjing Gong, Jennifer L Guerriero, et al. Best practices for spatial profiling for breast cancer research with the geomx® digital spatialprofiler. *Cancers*, 13(17):4456, 2021.

[17] Kok Hao Chen, Alistair N Boettiger, Jeffrey R Moffitt, Siyuan Wang, and Xiaowei Zhuang. Spatially resolved, highly multiplexed rna profiling in single cells. *Science*, 348(6233):aaa6090, 2015.

[18] Chee-Huat Linus Eng, Michael Lawson, Qian Zhu, Ruben Dries, Noushin Koulena, Yodai Takei, Jina Yun, Christopher Cronin, Christoph Karp, Guo-Cheng Yuan, et al. Transcriptome-scale super-resolved imaging in tissues by rna seqfish+. *Nature*, 568(7751):235–239, 2019.

[19] Xiao Wang, William E Allen, Matthew A Wright, Emily L Sylwestrak, Nikolay Samusik, Sam Vesuna, Kathryn Evans, Cindy Liu, Charu Ramakrishnan, Jia Liu, et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*, 361(6400):eaat5691, 2018.

[20] David Lahnemann, Johannes Koster, Ewa Szczurek, Davis J McCarthy, Stephanie C Hicks, Mark D Robinson, Catalina A Vallejos, Kieran R Campbell, Niko Beerenwinkel, Ahmed Mahfouz, et al. Eleven grand challenges in single-cell data science. *Genome biology*, 21(1):1–35, 2020.

[21] Duy Pham, Xiao Tan, Jun Xu, Laura F Grice, Pui Yeng Lam, Arti Raghubar, Jana Vukovic, Marc J Ruitenberg, and Quan Nguyen. Stlearn: integrating spatial

location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues. *BioRxiv*, 2020.

[22] Ruben Dries, Qian Zhu, Rui Dong, Chee-Huat Linus Eng, Huipeng Li, Kan Liu, Yuntian Fu, Tianxiao Zhao, Arpan Sarkar, Feng Bao, et al. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome biology,* 22(1):1–31, 2021.

[23] Jian Hu, Xiangjie Li, Kyle Coleman, Amelia Schroeder, Nan Ma, David J Irwin, Edward B Lee, Russell T Shinohara, and Mingyao Li. Spagcn: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nature method*, 18(11):1342–1351, 2021.

[24] Edward Zhao, Matthew R Stone, Xing Ren, Thomas Pulliam, Paul Nghiem, Jason H Bielas, and Raphael Gottardo. Bayesspace enables the robust characterization of spatial gene expression architecture in tissue sections at increased resolution. *bioRxiv*, 2020.

[25] Joseph Bergenstrahle, Ludvig Larsson, and Joakim Lundeberg. Seamless integration of image and molecular analysis for spatial transcriptomics workflows. *BMC genomics*, 21(1):1–7, 2020.

[26]https://www.10xgenomics.com/resources/datasets/human-breast-cancer-ductal-carcinoma-in-situ-invasive-carcinoma-ffpe-1-standard-1-3-0

[27] Dario Righelli, Lukas M Weber ,Helen L Crowell, Brenda Pardo, Leonardo Collado-Torres, Shila Ghazanfar, Aaron TL Lun, Stephanie C Hicks, and Davide Risso. Spatialexperiment: infrastructure for spatially-resolved transcriptomics data in r using bioconductor. *Bioinformatics*, 38(11):3128–3131, 2022.

[28] Davis J McCarthy, Kieran R Campbell, Aaron TL Lun, and Quin F Wills. Scater: pre-processing, quality control, normalization and visualization of single-cell rna-seq data in r. *Bioinformatics*, 33(8):1179–1186, 2017.

[29] Weber L, Crowell H (2022). ggspavis: Visualization functions for spatially resolved transcriptomics data. R package version 1.2.0, https://github.com/lmweber/ggspavis.

[30] Manuel Saiselet, Jo¨el Rodrigues-Vit´oria, Adrien Tourneur, Ligia Craciun, Alex Spinette, Denis Larsimont, Guy Andry, Joakim Lundeberg, Carine Maenhaut, and Vincent Detours. Transcriptional output, cell-type densities, and normalization in spatial transcriptomics. *Journal of molecular cell biology*, 12(11):906–908, 2020.

[31] Arian Lundberg, Linda S Lindstr¨om, J Chuck Harrell, Claudette Falato, Joseph W Carlson, Paul K Wright, Theodoros Foukakis, Charles M Perou, Kamila Czene, Jonas Bergh, et al. Gene expression signatures and immunohistochemical

subtypes add prognostic value to each other in breast cancer cohorts prognostic capacity of gene signatures, ki67, and routine ihc. Clinical Cancer Research, 23(24):7512–7520, 2017

[32] Zhen Yang, Andrew Wong, Diana Kuh, Dirk S Paul, Vardhman K Rakyan, R David Leslie, Shijie C Zheng, Martin Widschwendter, Stephan Beck, and Andrew E Teschendorff. Correlation of an epigenetic mitotic clock with cancer risk. Genome biology, 17(1):1–18, 2016.

[33] Scott L Carter, Aron C Eklund, Isaac S Kohane, Lyndsay N Harris, and Zoltan Szallasi. A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. Nature genetics, 38(9):1043–1048, 2006

[34] FM Buffa, AL Harris, CM West, and CJ Miller. Large meta-analysis of multiple cancers reveals a common, compact and highly prognostic hypoxia metagene. British journal of cancer, 102(2):428–435, 2010.

[35] Whijae Roh, Pei-Ling Chen, Alexandre Reuben, Christine N Spencer, Peter A Prieto, John P Miller, Vancheswaran Gopalakrishnan, Feng Wang, Zachary A Cooper, Sangeetha M Reddy, et al. Integrated molecular analysis of tumor biopsies on sequential ctla-4 and pd-1 blockade reveals markers of response and resistance. Science translational medicine, 9(379):eaah3560, 2017.