Computational analysis of super-resolved in situ sequencing data reveals genes modified by immune-tumor contact events

Michal Danino-Levi[1,2,3], Tal Goldberg[1,2,3], Maya Keter[1], Nikol Akselrod[1], Noa Shprach-Buaron[1,2,3], Modi Safra[1,2,3], Gonen Singer[1,*], Shahar Alon[1,2,3,*]

[1]The Alexander Kofkin Faculty of Engineering, [2]Gonda Multidisciplinary Brain Research Center, [3]Institute of Nanotechnology and Advanced Materials, Bar-Ilan University, Ramat Gan, 5290002, Israel

*Corresponding authors: shahar.alon@biu.ac.il; gonen.singer@biu.ac.il

Abstract

Cancer cells can manipulate immune cells and escape from the immune system response. Quantifying the molecular changes that occur when an immune cell is touching a tumor cell can increase our understanding of the underlying mechanisms. Recently, it became possible to perform such measurements in situ, for example using expansion sequencing, which enabled in situ sequencing of genes with super-resolution. We systematically examined whether individual immune cells from specific cell types express genes differently when in physical proximity to individual tumor cells. First, we demonstrated that a dense mapping of genes in situ can be utilized for the segmentation of cell bodies in 3D, thus improving our ability to detect likely touching cells. Next, we utilized three different computational approaches to detect the molecular changes that are triggered by proximity: differential expression analysis, tree-based machine learning classifiers, and matrix factorization analysis. This systematic analysis revealed tens of genes, in specific cell types, whose expression separates immune cells that are proximal to tumor cells from those that are not proximal, with a significant overlap between the different detection methods. Remarkably, an order of magnitude more genes are triggered by proximity to tumor cells in CD8 T cells compared to CD4 T cells, in line with the ability of CD8 T cells to directly bind Major Histocompatibility Complex (MHC) Class I on tumor cells. Thus, in situ sequencing of an individual biopsy can be used to detect genes likely involved in immune-tumor cell-cell interactions. The data used in this manuscript and the code of the InSituSeg, Machine learning, cNMF and Moran's I methods are publicly available at DOI: 10.5281/zenodo.7497981.

Introduction

The communication of the cancer cells with different types of cells that surround them, and in particular immune cells, can inhibit or promote tumor proliferation (Nishida-Aoki and Gujral 2019). Therefore, the study of cellular interactions within tumor tissues is essential for understanding the disease progression and the potential for its treatment (Wang, Lei and Han 2018). However, immune-tumor interactions in cancer tissue remain largely uncharacterized (Giladi *et al.* 2020). To obtain in depth characterization of immune-tumor cell-cell interactions, single cell quantification is needed. Alas, standard single-cell genomic technologies can profile each cell separately but only after tissue dissociation, therefore losing all information on cell locations in general, and on cell-cell interactions in particular. Single cell sequencing protocols can be modified to characterize immune-tumor interactions, for example using PIC-seq (Giladi *et al.* 2020). However, since this method uses small aggregates of cells, it is not trivial to reconstruct single-cell information, i.e., accurately assign the sequenced genes in each aggregate to their cell of origin.

A direct quantification of cell-cell interactions between individual immune and tumor cells can be obtained via in situ approaches, which utilize imaging to assess the identity and location of expressed genes. Spatially-resolved transcriptomics using technologies such as Slide-Seq and Spatial Transcriptomics (ST), allow sequencing of RNA fragments, potentially from all genes, to be mapped to their spatial location in human tissues and biopsies (Ståhl *et al.* 2016; Rodriques *et al.* 2019; Vickovic *et al.* 2019; Stickels *et al.* 2021). However, to date these technologies can't allow the detection and quantification of single cells. This is mainly because the tissue is dissolved in the process, prohibiting the acquisition of cellular morphological features such as DAPI staining for the nucleus, and also because the resolution is not high enough for single cell analysis. Although there are computational attempts to reconstruct single-cell information from this data (Elosua-Bayes *et al.* 2021; Rao *et al.* 2021), and to integrate this data with single cell sequencing (Kleshchevnikov *et al.* 2020; Longo *et al.* 2021; Cable *et al.* 2022), accurate assignment of genes to single cells is still a challenge. Technologies based on multiplexed fluorescent in situ hybridization (FISH) allow measuring tens and even hundreds of genes in situ with a single cell resolution. These technologies include MERFISH (Moffitt *et al.* 2016) as well as SeqFISH, STARmap, ISS, RNAscope, BOLORAMIS and more (Ke *et al.* 2013; Codeluppi *et al.* 2018; Wang *et al.* 2018; Eng *et al.* 2019; Liu *et al.* 2021). A recent technology, termed expansion sequencing or ExSeq, allows in situ sequencing with super-resolution (Alon *et al.* 2021). Here we utilize an ExSeq measurement of 297 genes in a human breast cancer biopsy to perform a new kind of analysis – quantification of gene expression modifications in single interacting cells in situ. We identify physically touching cells with super-resolution, quantify immune-tumor cell-cell interactions, and determine how an immune cell is changing its gene expression profile when it is close to a tumor cell, and vice versa (Fig. 1).

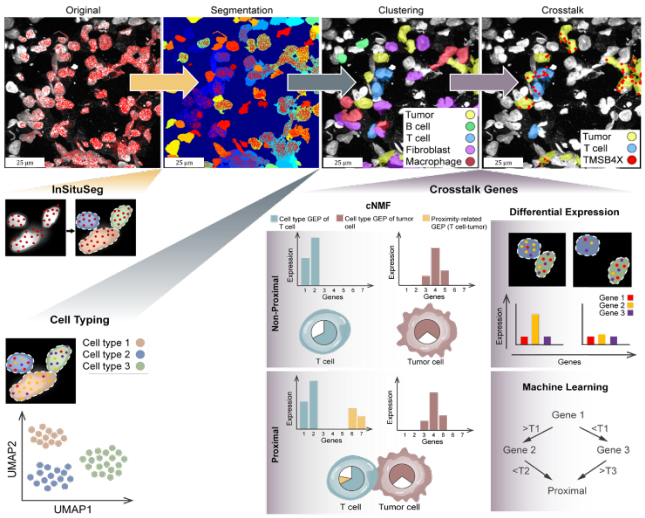

Figure 1. Overview of the detection of immune-tumor crosstalk genes. First, the ExSeq images were segmented using InSituSeg. Next, cell typing was performed using the cell's expression profiles, clustered after dimension reduction and displayed via Uniform Manifold Approximation and Projection (UMAP). Finally, crosstalk genes were detected using a differential expression, tree-based machine learning methods, and matrix factorization using cNMF (Kotliar *et al.* 2019). In the cNMF panel, Gene Expression Profile (GEP) can define cell type (blue=T cells, brown=tumor cells), or be proximity-related (yellow). In the schema, two GEPs represent cell types, and one GEP is triggered by proximity. The pie chart inside each cell describes its GEP usage.

Materials and Methods

Description of the datasets

Biopsies were collected from patients at Dana Farber Cancer Institute and originally described in (Alon *et al.* 2021). The sample utilized in this study was of a liver metastasis of hormone receptor positive breast cancer. The region sequenced in situ with ExSeq was 1347 x 621 x 8 microns in size (before expansion). Full description of biopsy and the 297 interrogated genes is in the supplementary information.

Segmentation of cell bodies

We developed a segmentation pipeline, termed InSituSeg, that takes advantage of the dense mapping of genes in situ for segmentation of cell bodies in 3D, using staining of cell nuclei and RNA locations (Fig. 2A). The steps of the segmentation pipeline (Fig. 2B) and its parameters (Table S2) are described in the supplementary information.

Clustering segmented cells

In order to identify and cluster the segmented cells according to their expression pattern, we utilized the R toolkit Seurat (Hao *et al.* 2021), and followed the analysis in (Alon *et al.* 2021). The procedure is described in the supplementary information.

Detecting differentially expressed genes

For any pair of cell clusters X and Y, cluster X was partitioned into two subsets: a subset of X cells that are proximal to Y cells, and a subset of X cells that are not proximal to Y cells. Comparisons were performed to observe differences in non-tumor cell types when in proximity to tumor cell types, and vice versa. Gene expression change (fold change) and p-value per gene in each comparison were calculated using DESeq2 (Love, Huber and Anders 2014), and we proceeded with genes that had Benjamini-Hochberg false discovery rate (FDR) of 0.1. To further assess the statistical significance of the results, we used permutation analysis. To avoid errors that result from inaccurate boundaries detection of two adjacent cells, we filtered genes in two different ways: 1) We filtered upregulated genes detected in X cells if they are known cell markers for the Y cells (the known marker genes are listed in the Methods section 'Description of the datasets'). 2) We filtered genes detected in X cells (i.e., induced in the subset of X cells that are proximal to Y cells compared to the subset of X cells which are not proximal) if they are highly differentially expressed in the Y cells (i.e., induced in the subset of Y cells that are proximal to X cells compared to the subset of Y cells which are not proximal). High degree of overlap exists between the two different filtering methods, full details are in supplementary information.

Machine learning pipeline

We applied machine learning tools to detect genes that their expressions separate, for cell type X, cells that are proximal to cell type Y versus non proximal cells. In contrast to the detection of differentially expressed genes described above, machine learning tools can detect genes that change their expression in concert due to the proximity between cells. Overall four machine learning classifiers were applied on the dataset: Decision Trees (Quinlan 1986), Random Forest (Ho 2002), XGBoost (Chen and Guestrin 2016) and CatBoost (Dorogush, Ershov and Gulin 2018). To evaluate the performance of the classifiers, we first checked how sensitive the results are with respect to the initial (random) decision of which part of the dataset will serve as a train and which part will be the test. Then we compared the results obtained to the results of the same dataset, but with the class labels shuffled such that it should not contain biological meaning. To avoid errors that result from inaccurate boundaries detection of two adjacent cells, we filtered upregulated genes detected in X cells if they are known cell markers for the Y cells (the known marker genes are listed in the supplementary information, section 'Description of the datasets'). Full details are in supplementary information.

cNMF analysis

We implemented cNMF (Kotliar *et al.* 2019) analysis with the aim of detecting a battery of genes that change their expression together as a result of proximity between immune and tumor cells. Using this analysis we discovered gene signatures, namely gene expression programs ('GEP), which define cell types as well as cell states. We examined whether

GEPs can be overexpressed or under expressed in a cell type as a result of physical distance from other cell types ('proximity-related' GEPs). Full details are in supplementary information.

Quantifying the statistical significance of overlapping genes

We assessed the statistical significance of the overlapping genes between any two detection methods (differential expression, machine learning and matrix factorization) with a bootstrapping approach (supplementary information).

Moran's I calculation

We implemented Moran's I calculation in the context of spatially-resolved transcriptomics. We compute a p-value for the spatial dependence of each gene by taking into account the locations of all the genes expressed in the tissue. P-values were estimated using bootstrapping. We also implemented Moran's I on the level of cells from a given cell type. I.e., For each cell type we generate a p-value for the spatial dependence of the cells in the given type. Full details are in supplementary information.

Results and Discussion

3D segmentation of single cells bodies using in situ sequencing data

With the aim to characterize immune-tumor cell-cell interactions, we utilized a spatial dataset of a core biopsy, 1347 x 621 x 8 microns in size, taken from a patient with metastatic breast cancer infiltration into the liver, and sequenced in situ via targeted ExSeq (Alon *et al.* 2021). With targeted ExSeq, a set of genes is selected and then oligonucleotide padlock probes bearing barcodes for each selected gene are hybridized to specific transcripts. These padlock probes are amplified in situ to generate amplicons for subsequent readout through in situ sequencing of the barcodes. The resulting sequenced amplicons (termed reads) give the precise location of the transcripts in situ. In this biopsy, 297 genes were characterized in situ with super resolution (Table S1), due to 3.3x physical expansion of the tissue. The interrogated genes included gene markers for cell types and genes known or suspected to be associated with cancer tissues (Methods).

We developed a pipeline for ascribing in situ sequencing reads to cell bodies, termed InSituSeg, which aids in pinpointing touching cells in 3D, even in a densely packed tumor tissue (Fig. S1-2). The main idea of InSituSeg is to utilize the dense mapping of genes in situ for the segmentation of cell bodies in 3D (Fig. 2). Segmentation of cells is typically performed using only nuclei staining, without using information about RNA location (Stringer *et al.* 2021; Hollandi *et al.* 2022). This procedure doesn't maximize the number of sequencing reads assigned to cells, mainly because sequencing reads are often located in the cell soma outside the nucleus. In contrast to recent tools (Hu *et al.* 2021; Littman *et al.* 2021; Petukhov *et al.* 2022), InSituSeg doesn't use prior information about cell types, or even information about RNA identities (i.e., genes). Instead, it uses only imaging data resulting from a typical in situ sequencing experiment: DAPI staining and the locations of the sequenced RNA molecules (Fig. 2, Fig. S3-4 and Methods). InSituSeg is performed in 3D, which aids in separation of cells which seem to be overlapping when looking only at the x-y plane (Fig. S1); and therefore has advantage compared to 2D watershed-based segmentation which is performed on individual z-planes. Importantly, since cell type information is not used, InSituSeg can ascribe an atypical gene to a cell from a given cell type, and thus can possibly better represent the heterogeneity of individual cells.

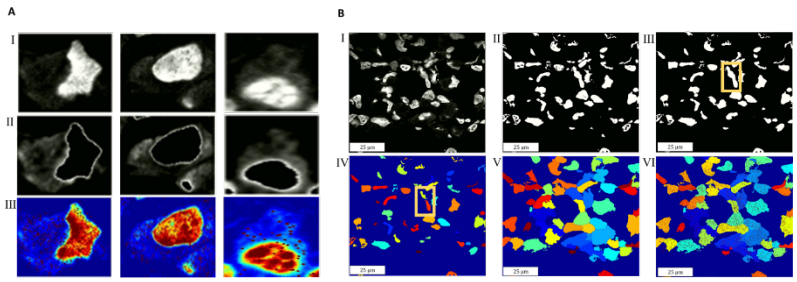

Figure 2. Scheme of the InSituSeg pipeline. The pipeline utilizes dense mapping of genes in situ for segmentation of cell bodies, using pixel intensity thresholding of 3D images. The input is a DAPI stained image and the spatial locations of the mRNA molecules, and the output is a segmented 3D image with the mRNA assigned to each cell body. (A) Three cells are presented (columns). I) After DAPI staining, the signal of the cell bodies is weaker compared to the strong nucleus staining of genomic DNA, but nevertheless can be clearly detected in the examined cells with pixel intensity thresholding (II). III) A clear overlap between the hues observed in the DAPI image and the sequenced RNA (red dots, the DAPI intensities are shown in red-blue for better visualization). This overlap further confirms that the hues correspond to cell bodies. (B) The segmentation pipeline is composed of six steps (Methods): I) Illumination correction. II) Detection of nuclei voxels. III) Refinement of nuclei voxels. IV) Splitting of large nuclei. For example, the large putative nucleus marked by a yellow rectangle in (III) is split in (IV) into two nuclei. V) Detection of cell body voxels using watershed segmentation. VI) Assignment of mRNA molecules into cell bodies.

InSituSeg utilizes pixel intensity thresholding to reduce the strong nuclei staining of the DAPI and reveal residual DAPI staining in the cytosol (Fig. 2 and Methods). Residual DAPI staining in the cytosol was demonstrated before in the context of multiplexed FISH imaging, and was termed 'cytoDAPI' (Wang *et al.* 2021). This residual DNA staining can be a result of RNA staining (as suggested in (Wang *et al.* 2021)) or due to staining of rolonies (i.e., the padlocks which bind single molecule RNA, after phi29 amplification). Rolonies might be double stranded to some extent due to limited template switching of phi29 (Ducani, Bernardinelli and Högberg 2014). Residual DAPI staining might also be influenced by cytoplasmic DNA which is more prevalent in tumor cells (Anindya 2022). However, in our data tumor cells don't have on average larger residual DAPI staining in the cytosol compared to other cell types (Fig. S5). Details about the parameters used by InSituSeg and the sensitivity to fine tuning them are provided (Methods and Fig. S6).

To test the performance of InSituSeg, we: a) showed that it is in agreement with manual segmentation (Fig. S7A); b) tested it on a different core biopsy that was analyzed by expansion sequencing (Fig. S7B); c) demonstrated that it outperforms two recent neuronal network-based segmentation tools, ilastik (Berg *et al.* 2019) and Mesmer (Greenwald *et al.* 2022) (Fig. S7C). We note that in contrast to ilastik, Mesmer and other recent segmentation tools, InSituSeg is specifically designed for in situ sequencing image data, and therefore InSituSeg is not a general purpose segmentation tool; d) showed that InSituSeg is superior to using RNA sequencing data alone for grouping reads into cells, i.e. without using nuclei information (Fig. S7C and Methods); e) demonstrated that InSituSeg can be applied for in situ imaging data generated with MERFISH (Fig. S8); and finally, f) estimated that InSituSeg captures between 65 to 71% of the cell body area, as determined via cytosolic and membrane staining (Fig. S9). We further estimated that segmentation with InSituSeg can add 20% to the cell body volume, compared to nuclei segmentation alone (Fig. S9).

Overall, the dataset of the core biopsy contained 1,146,615 spatially resolved sequenced reads from 297 genes. Manual segmentation of nuclei using the tool VAST (Berger, Seung and Lichtman 2018) resulted in 2,395 cells (reporting only cells with at least 100 reads per cell), and 771,904 reads were assigned to them (Alon *et al.* 2021). In contrast, using InSituSeg, 2,748 cells were detected, and 939,764 reads were assigned to them (again only cells containing at least 100 reads are reported). Thus, InSituSeg gives a 15% and 22% increase in the number of segmented cells and the number of assigned reads, respectively (Fig. S10A), which can lead to better characterization of the molecular content of the cells.

Moreover, the detection of cell bodies with InSituSeg, combined with the super-resolution of ExSeq, allowed pinpointing touching cells in 3D below (Fig. 1 segmentation step and Fig. S1-2).

We next performed expression clustering on the InSituSeg results and compared it to manual segmentation of nuclei (Alon *et al.* 2021) (Fig. S10B-C). The analysis was done using principal component analysis (PCA)-based expression clustering of Seurat (Hao *et al.* 2021), and displayed using the Uniform manifold approximation and projection (UMAP) representation (Becht *et al.* 2018) (Methods). Overall, in both approaches, the expression clustering revealed the expected mixture of cell types, including tumor, immune (T cell, B cell, and macrophage), and fibroblast cell clusters (Fig. S10B-C). However, with InSituSeg the higher number of reads assigned to cells allowed us to classify an additional tumor subtype, marked by the gene EPCAM. Finally, the transcriptionally-defined cell clusters were mapped onto tissue context (Fig. S10D-E).

Identification of genes involved in cell-cell interactions using differential expression

We next utilized the ExSeq data, after processing with InSituSeg and expression clustering, to characterize immune-tumor cell-cell interactions in situ. Specifically, we aimed to detect genes in a given cell type that have different expressions as a result of proximity to another cell type. These genes can either be influenced by the proximity between the cells, or even influence the proximity to occur. We first utilized a differential expression approach (Methods): For any pair of cell clusters X and Y, cluster X was partitioned into two subsets: a subset of X cells that are proximal to Y cells (i), and a subset of X cells that are not proximal to Y cells (ii), and all differentially expressed genes between (i) and (ii) were detected using DeSeq2 (Love, Huber and Anders 2014). The resulting p-values were further validated using bootstrapping (Methods). Cell-cell proximity was estimated using the smallest Euclidean distance between the mRNA molecules in two adjacent cells, utilizing the InSituSeg cell body segmentation. We set a threshold of 3 microns (before expansion) for that distance, and validated the robustness of the results to changes in this parameter (Fig. S11-12). Taking advantage of the super-resolution, which is a result of the physical expansion in ExSeq, we examined distances between cell bodies down to half a micron (Fig. S12), further increasing the likelihood that the cells are physically touching. The genes detected below as induced by proximity are consistent between the different distance cutoffs (Fig. S12, Table S3). The physical expansion of ExSeq also allows a large number of transcripts to be quantified together, since neighboring RNA molecules can be better resolved (Xia *et al.* 2019; Alon *et al.* 2021). We estimate that without expansion most amplified transcripts would not be resolved due to spatial overlap (Fig. S13 and Methods). The dramatic decrease in the number of amplified transcripts resolved would have been also manifested in a decrease in the proximity-induced genes that can be detected (Fig. S13E and Methods).

We systematically examined all possible interactions between tumor (5 cell clusters, Fig. S10C) and non-tumor cell types (7 cell clusters, Fig. S10C), overall 108 comparisons (Methods). We accounted for multiple testing using a Benjamini-Hochberg false discovery rate (FDR) of 0.1. The systematic search resulted in 11.8 genes, on average, detected as differentially expressed in the 108 comparisons performed (Fig. 3, Fig. S14 and Fig. S15). Note that with bootstrapping, which is utilized to compute the p-values (Methods), on average less than one gene was detected as proximity-induced. This is true for the original cutoff distance of 3 microns, as well as all the other cutoff distances down to 0.5 microns. An example of proximity-induced gene is the gene thymosin beta 4 X-Linked (TMSB4X), which is involved in the organization of the cytoskeleton, is overexpressed when CD3D-positive T cells are in proximity to tumor cells in general, compared to

CD3D-positive T cells which are not proximal (Fig. S14). This gene is also upregulated when T cells in general are proximal to EPCAM-positive tumor cells, when CD8A-positive T cells are proximal to tumor cells in general, and also when T cells in general are proximal to CD44-positive tumor cells (Fig. S14). Consequently, comparing all T cells that are proximal to tumor cells in general, to T cells that are not proximal, also reveals that this gene is overexpressed (Fig. 3A and S14). Interestingly, in the last few years, this gene was detected as upregulated in breast cancer, and it was suggested that its expression correlates with poor prognosis (Zhang *et al.* 2017; Morita and Hayashi 2018). The data presented here might point to the exact settings in which this gene is upregulated.

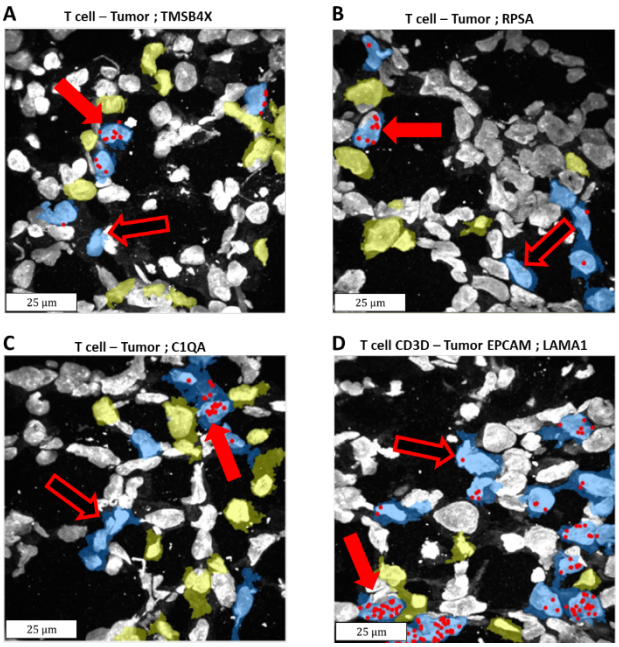

Figure 3. Example of genes identified as induced in T cells when proximal to tumor cells. Sequencing reads locations (red spots) of four induced genes are overlaid on the DAPI staining of the nuclei, as well as the segmentation of T cells (blue) and tumor cell types (yellow). The cell bodies were detected using InSituSeg, and the cell types were identified using clustering of the gene expression profiles. Only segmentations of T cells and tumor cells are presented. Genes upregulated in T cells due to proximity to tumor cells have more red spots when proximal to tumor cells (exemplars in full red arrows versus hollow red arrows). A) the gene Thymosin Beta 4 X-Linked (TMSB4X) was detected by differential expression (DE), by matrix factorization (MF), and by machine learning (ML), when examining all T cells and all tumor cells. B) The gene Ribosomal Protein SA (RPSA) was detected by DE, by ML, and by MF, when examining all T cells and all tumor cells. C) the gene Complement Component 1, Q Subcomponent, A Chain (C1QA) was detected by DE when examining all T cells and all tumor cells. D) the gene Laminin Subunit Alpha 1 (LAMA1) was detected by DE and by ML, when examining the subtype T cell-CD3D and the subtype tumor-EPCAM. Each panel shows a subset region from the biopsy, acquired with a 40X objective, 100 x 100 microns in size (before expansion). Note that max projection is shown and therefore some cells seem to overlap, but they are clearly separated in 3D (Fig. S1). DE was performed with DeSeq2 (Love, Huber and Anders 2014), ML with CatBoost (Dorogush, Ershov and Gulin 2018), and MF with cNMF (Kotliar *et al.* 2019). Permutation analysis was performed on all methods to assess statistical significance.

Identification of genes involved in cell-cell interactions using machine learning tools

We then applied supervised machine learning tools to identify genes that their expression separates, for cell type X, cells that are proximal to cell type Y versus non proximal cells. We focused on Decision Tree (Quinlan 1986), a classifier with a high level of interpretability, and on algorithms that are based on Decision Trees with a low level of interpretability, including Random Forest (Ho 2002), XGBoost (Chen and Guestrin 2016) and CatBoost (Dorogush, Ershov and Gulin 2018). We designed and applied a machine learning pipeline (Fig. S16 and Methods) on each one of the 108 comparisons between non-tumor cell types and tumor cell types as described above, using the same measure of physical proximity between cells (Methods and sensitivity test in Fig. S17). The data for each comparison, i.e., the gene expression of the cells that can either be in proximity or not-proximal to the different cell type, was randomly split into training and testing sets. The split was stratified so the relative distribution of the proximity vs not-proximal cells was retained. The testing set was not used during the training phase, and on the training set we applied the stratified k-fold cross validation strategy (Tan and Gilbert 2003). In most cases, the number of non-proximal cells was higher than proximal cells, therefore the dataset was

imbalanced and we utilized over-sampling methods to correct this effect (Methods). We ran multiple combinations of the classifiers' hyperparameters to find the best ones for each classification algorithm ('best model', Methods). For each comparison, we determined the classifier with the best performance ('best classifier'), which was then applied to the test set.

When detecting genes that classify cells as proximal and non-proximal, the results are expected to be more robust when the number of cells, both proximal and not-proximal, is high. On the other hand, when studying a biopsy from an individual patient using spatially-resolved transcriptomics, the overall number of cells studied is typically on the order of thousands (Fig. S10E) or tens of thousands. In addition, refining the cell types in the comparison, for example studying subtypes of T cells and subtypes of tumor cells, is expected to produce more specific results, but reduce the number of cells even further. Therefore, the number of cells fed to the classifier was not high overall (Tables S4-5). Specifically, in most comparisons, the number of proximal cells per cross validation fold was ~10-20 (or the number of non-proximal cells in cases when their number is lower than proximal cells, Tables S4-5), making the detection of proximity-induced genes challenging with machine learning tools. Therefore, we: a) quantified the sensitivity of the results with respect to the initial random split between train and test data (Methods); this sensitivity is expected to be high due to the small number of cells. b) performed an additional evaluation of the performance of the classifiers using non-biological realizations. These realizations were generated by using the same dataset, but with the cell labels (proximity or not-proximal) shuffled such that they should not contain biological meaning (Methods). For the best classifier, we generated 30 non-biological realizations for each comparison, and for each realization, the machine learning pipeline was re-run. We compared the results to 30 runs of the pipeline with the best classifier using the original (unshuffled) data, each run with a different initial random split between train and test data, and computed bootstrap p-values. We kept only the machine learning results that had Benjamini-Hochberg false discovery rate (FDR) less than 1e-4. Finally, for comparisons that passed the aforementioned test, the best classifiers were applied to the complete dataset, i.e., without splitting into train and test (Methods).

Interestingly, the CatBoost classification method was found to outperform the three other classification methods in all 108 comparisons. Overall, only 60 out of the 108 comparisons resulted in classifications that were deemed significant with FDR<1e-4 (Tables S4-5). The top ten features (i.e. genes) that give rise to the significant classifications are presented in Fig. S18. The detected genes can either be upregulated or downregulated due to the proximity between cells from different types (Fig. S18). Note that errors that result from inaccurate boundaries detection of two adjacent cells, or mis-segmentation, can lead to cases of false detection of proximity-induced genes. We filtered genes which were likely to be detected due to inaccurate boundaries detection using two approaches (see Methods). However, some cases of mis-segmentation-based errors might still occur. For example, in Fig. S18 the gene Pecam1 (CD31) appears as proximity-induced in B cells when close to tumor-ALDH1A3. While Pecam1 can be expressed in B cells, it is known to be expressed in endothelial cells, and therefore mis-segmentation of blood vessels might have contributed to this result. Likewise, while LYZ might be expressed in tumor cells in breast cancer (Vizoso *et al.* 2001), it is known to be expressed in immune cells, and therefore the detection of this gene as proximity-induced in Tumor-PGR when close to T cells (Fig. S18), might be due to mis-segmentation.

The differential expression approach and the machine learning classifiers are fundamentally different, so we didn't expect a one-to-one agreement between the genes detected using both approaches. Nevertheless, many genes did overlap; comparing non-tumor cell types that are proximal to tumor cell types versus non-tumor cell types that are not proximal to

tumor cells, out of 436 genes detected using a differential expression, 61 genes were also detected using machine learning (Table S6). The overlap is even more profound when examining the other direction, i.e., tumor cell types that are proximal to non-tumor cell types, versus tumor cell types that are not proximal to non-tumor cells; out of 840 genes detected using a differential expression, 166 genes were also detected using machine learning (Table S6). Overall, in 56 out of the 108 comparisons performed between all tumor and immune cell types, genes were detected as proximity-related using both the differential expression approach and machine learning. Importantly, the overlap between the detected genes was statistically significant (p-value<0.05, bootstrapping, Methods) in 37 out of these 56 comparisons (Table S6). This overlap between the approaches provides additional support for the validity of the findings. Examining the genes detected by both differential expression and machine learning, taking T cells for example, clearly show the overexpression of these genes when proximal to tumor cells (Fig. 4).

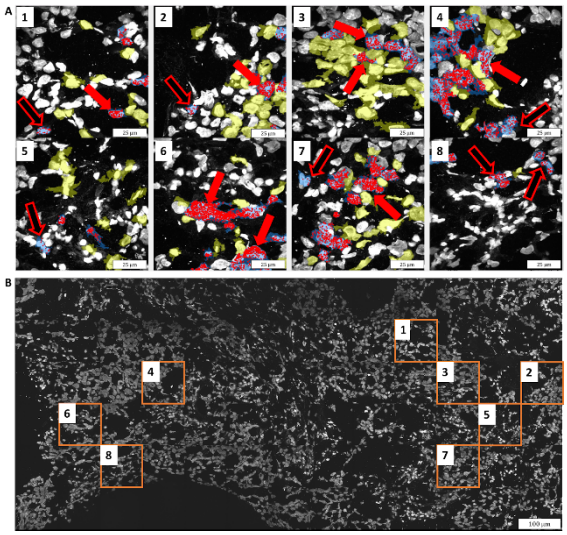

Figure 4. Overexpression of a group of genes in T cells when proximal to tumor cells.

Six genes were detected as induced by both differential expression and machine learning when T cells are proximal to tumor cells: RPSA, CD63, LYZ, TMSB4X, S100A14, LAMA1. A) Sequencing reads locations (red spots) of these genes are overlaid on the DAPI staining of the nuclei, as well as the segmentation of T cells in blue and tumor cell types in yellow. The cell bodies were detected using InSituSeg, and the cell types were identified using clustering of the gene expression profiles. Only segmentations of T cells and tumor cells are presented. Genes upregulated in T cells due to proximity to tumor cells have more red spots (overexpression) when proximal to tumor cells (exemplars in full red arrows versus hollow red arrows). B) The biopsy with DAPI staining. Each panel in (A) shows a max projection of a subset region from the biopsy (orange square in (B)), acquired with a 40X objective, 100 x 100 microns in size (before expansion).

An example of a gene detected using both differential expression analysis and machine learning is Keratin 19 (KRT19). KRT19 was detected using differential expression analysis as upregulated in tumor cells proximal to T cells, compared to tumor cells not proximal to T cells (Fig. S14). This gene was also detected using machine learning as the highest classification feature for all tumor cells proximal to all T cells, versus tumor cells not proximal (Fig. S18). This gene is also the second highest classification feature for all tumor cells proximal to CD8A-positive T cells versus not proximal tumor cells, and the second highest classification feature for EGFR-positive tumor cells proximal to CD8A-positive T cells, versus not proximal EGFR-positive tumor cells (Fig. S18). KRT19 is known to be important for the structural integrity of epithelial cells, and is a marker gene for breast tumors (Saha *et al.* 2017). Our analysis pinpoints the settings in which this gene is upregulated, namely that this gene expression might be higher when tumor cells are proximal to T cells, and in particular to CD8A-positive T cells.

Remarkably, a clear difference is observed between CD4 and CD8 T cells, in line with the ability of CD8 T cell to directly bind Major Histocompatibility Complex (MHC) Class I on tumor cells. 12 and 10 genes were detected as overexpressed when CD8-positive T cells are in proximity to tumor cells, using differential expression and machine learning analysis, respectively. In contrast, only 1 and 0 such genes were detected when CD4-positive T cells are in proximity to tumor cells, using differential expression and machine learning analysis, respectively (Table S6). Likewise, 38 and 10 genes were detected as overexpressed when tumor cells are in proximity to CD8-positive T cells, using differential expression and

machine learning analysis, respectively. In contrast, only 6 and 0 such genes were detected when tumor cells are in proximity to CD4-positive T cells, using differential expression and machine learning analysis, respectively (Table S6). Thus, while in CD8 T cells physical proximity to tumor cells trigger changes in gene expression in both the T cell and the tumor cell, in CD4 T cells the changes in genes expression might be more gradual with respect to the distances to tumor cells.

Identification of genes involved in cell-cell interactions using matrix factorization

We then applied matrix factorization to identify a battery of genes that change their expression together as a result of proximity between immune and tumor cells (Fig. S19). We utilized cNMF (Kotliar *et al.* 2019) to discover gene signatures, termed Gene Expression Programs (GEP), which define cell types as well as cell states. Specifically, we examined whether GEPs can be proximity-related, i.e., can be overexpressed or under expressed in a cell type as a result of physical distance from other cell types. To do so we divided each non-tumor cell type into two subgroups: cells proximal to tumor cells versus cells that are not close to tumor cells (Methods). A similar analysis was performed in the other direction (i.e., tumor cells proximal or not proximal to immune cells). Then, for each GEP in each cell type, we compared the usage of that GEP in the proximal cells subgroup to the usage in the non-proximal subgroup, and computed statistical significance using permutation analysis (Methods). Importantly, this analysis revealed 6 GEP which are induced by proximity to tumor cells (Fig. S19, Table S7). Importantly, in one such proximity-related GEP, expressed in T cells, 8 out of the 15 genes in this GEP overlapped with the genes detected using differential expression (significant overlap, p-value<0.01, bootstrapping, Tables S6-7-8). In addition, 5 out of the 15 genes in this GEP overlapped with the genes detected using machine learning (significant overlap, p-value<0.01, bootstrapping, Tables S6-7-8). The overlaps found between these three computational approaches further support the possibility that the detected genes are indeed involved in cell-cell interactions between immune and tumor cell types (Fig. 3-4). Thus, the detected genes can potentially be markers for immune reactions toward tumor cells, or vice versa.

Detection of proximity-induced genes as a function of the fraction of data utilized

We explored the dependency of the number of proximity-induced genes on the fraction of the data used, and the number of adjacent non-tumor cells to tumor cells, via a scale-down analysis (Methods). This analysis revealed a linear trend between the fraction of the data utilized and the number of proximity-induced genes revealed in T cells (Fig. S20). Importantly, a linear trend is also observed between the number of proximity-induced genes in T cells and the number of adjacent T cells and tumor cells (Fig. S20). The linear trend is also evident in other non-tumor cell types (Fig. S21). This trend suggests that a rational design of experiments aimed at detecting proximity-induced genes is feasible, given the number of adjacent cells present in the studied biopsy.

Detecting spatially-dependent genes and cell types

We next examined if the genes detected as involved in cell-cell interactions tend to be spatially-dependent. For this, we implemented Moran's I measurement for segmentation-free detection of spatially-dependent genes (Hao *et al.* 2021; Hu *et al.* 2021). Similarly to a recent implementation (Miller *et al.* 2021), we account for non-uniform cell distribution in the tissue. Our implementation detects specific genes that have higher spatial dependence, relative to other expressed genes,

by using the distribution of locations of all the genes in the tissue (Methods). For Moran's I calculation we automatically select the grid (spatial bins) that produces the most robust results for the spatially-dependent genes (Methods).

Overall, 169 genes were detected as spatially-dependent (FDR<0.01), out of the 297 genes interrogated. Note that the selection of genes in the ExSeq gene panel potentially explains the large fraction of spatially variable genes. Ranking the genes according to their p-value, arranged from the smallest to the largest, we examined the top detected genes and six of them are presented in Fig. S22: KIT, S100A8, SOX18, COBL, RPSA, and XBP1. RPSA, Ribosomal Protein SA, was detected as regulated by proximity between T cells and tumor cells in the differential expression analysis, the machine learning analysis, and the cNMF analysis (Fig. 3B, Fig. S14 and S18, Table S6). The expression of RPSA is increased in many cancers including breast, and clinical trials are ongoing to test if it can serve as a biomarker of tumor invasion in pancreatic ductal adenocarcinoma (clinical trials identifier (NCT number): NCT04575363). The spatial dependence of this gene (Fig. S22E), as well as the possible upregulation of this gene due to T cells and tumor cells proximity (Fig. S14 and S18), suggest that this gene might serve as a biomarker in breast cancer as well. However, given that most of the examined genes were detected as spatially-dependent, it is unlikely that the main cause for the spatial dependence is the involvement in cell-cell interactions. We note that the spatial dependence of the genes can't be fully explained by uneven spatial distribution of cell types, since we detected genes that are spatially variable in spite (or in excess) of cell type spatial variability (Methods, Fig. S23 and Table. S9). Manual examination of the data did not reveal a clear link between the locations of genes that are spatially variable in excess of their cell type and the locations of the potentially interacting cell types (Fig. S24). Lastly, we also examined the spatial dependence of the cell types, revealing a clear difference between non-tumor cell types versus tumor cells (Fig. S25-26). Segmentation-free detection opens the door to the analysis of several genes and cell types that might be interacting in specific locations in the tissue.

Data Deposition

The data used in this manuscript and the code of the InSituSeg, Machine learning, cNMF and Moran's I methods are publicly available at DOI: 10.5281/zenodo.7497981. Test data for running the code is also provided in the same deposit.

Acknowledgments

This work was supported by the Israel Science Foundation (ISF) [grant numbers 2958/21, 3363/21]; Israel Cancer Association (ICA) [grant number 20220069]; Israel Ministry of Science [grant number 2180]; Brightfocus Foundation [grant number 929965]; Joint Sheba-Bar Ilan Research Grant; and European Research Council (ERC) (Grant number: 101117324). We would like to thank Lion Morgenstein for contributing analysis for Fig. S13.

Conflict of Interest: none declared.

References

Alon S, Goodwin DR, Sinha A *et al.* Expansion sequencing: Spatially precise in situ transcriptomics in intact biological systems. *Science* 2021;**371**:eaax2656.

Anindya R. Cytoplasmic DNA in cancer cells: Several pathways that potentially limit DNase2 and TREX1 activities. *Biochim Biophys Acta Mol Cell Res* 2022;**1869**:119278.

Becht E, McInnes L, Healy J *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol* 2018;**37**:38–44.

Berg S, Kutra D, Kroeger T *et al.* Ilastik: Interactive machine learning for (bio)image analysis. *Nat Methods* 2019;**16**:1226–32.

Berger DR, Seung HS, Lichtman JW. VAST (Volume Annotation and Segmentation Tool): Efficient manual and semi-automatic labeling of large 3D image stacks. *Front Neural Circuits* 2018;**12**:88.

Cable DM, Murray E, Zou LS *et al.* Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat Biotechnol* 2022;**40**:517–26.

Chen T, Guestrin C. XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 2016, DOI: 10.1145/2939672.2939785.

Codeluppi S, Borm LE, Zeisel A *et al.* Spatial organization of the somatosensory cortex revealed by osmFISH. *Nat Methods* 2018;**15**:932–5.

Dorogush AV, Ershov V, Gulin A. CatBoost: gradient boosting with categorical features support. 2018, DOI: 10.48550/ARXIV.1810.11363.

Ducani C, Bernardinelli G, Högberg B. Rolling circle replication requires single-stranded DNA binding protein to avoid termination and production of double-stranded DNA. *Nucleic Acids Res* 2014;**42**:10596–604.

Elosua-Bayes M, Nieto P, Mereu E *et al.* SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic Acids Res* 2021;**49**:e50.

Eng C-HL, Lawson M, Zhu Q *et al.* Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature* 2019;**568**:235–9.

Giladi A, Cohen M, Medaglia C *et al.* Dissecting cellular crosstalk by sequencing physically interacting cells. *Nat Biotechnol* 2020;**38**:629–37.

Greenwald NF, Miller G, Moen E *et al.* Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nat Biotechnol* 2022;**40**:555–65.

Hao Y, Hao S, Andersen-Nissen E *et al.* Integrated analysis of multimodal single-cell data. *Cell* 2021;**184**:3573-3587.e29.

Ho TK. Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*. IEEE Comput. Soc. Press, 2002.

Hollandi R, Moshkov N, Paavolainen L *et al.* Nucleus segmentation: towards automated solutions. *Trends Cell Biol* 2022;**32**:295–310.

Hu J, Li X, Coleman K *et al.* SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat Methods* 2021;**18**:1342–51.

Ke R, Mignardi M, Pacureanu A *et al.* In situ sequencing for RNA analysis in preserved tissue and cells. *Nat Methods* 2013;**10**:857–60.

Kleshchevnikov V, Shmatko A, Dann E *et al.* Comprehensive mapping of tissue cell architecture via integrated single cell and spatial transcriptomics. *bioRxiv* 2020, DOI: 10.1101/2020.11.15.378125.

Kotliar D, Veres A, Nagy MA *et al.* Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *Elife* 2019;**8**, DOI: 10.7554/eLife.43803.

Littman R, Hemminger Z, Foreman R *et al.* Joint cell segmentation and cell type annotation for spatial transcriptomics. *Mol Syst Biol* 2021;**17**:e10108.

Liu S, Punthambaker S, Iyer EPR *et al.* Barcoded oligonucleotides ligated on RNA amplified for multiplexed and parallel in situ analyses. *Nucleic Acids Res* 2021;**49**:e58.

Longo SK, Guo MG, Ji AL *et al.* Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. *Nat Rev Genet* 2021;**22**:627–44.

Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:550.

Miller BF, Bambah-Mukku D, Dulac C *et al.* Characterizing spatial gene expression heterogeneity in spatially resolved single-cell transcriptomic data with nonuniform cellular densities. *Genome Res* 2021;**31**:1843–55.

Moffitt JR, Hao J, Wang G *et al.* High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proc Natl Acad Sci U S A* 2016;**113**:11046–51.

Morita T, Hayashi K. Tumor progression is mediated by thymosin-ß4 through a TGFß/MRTF signaling axis. *Mol Cancer Res* 2018;**16**:880–93.

Nishida-Aoki N, Gujral TS. Emerging approaches to study cell-cell interactions in tumor microenvironment. *Oncotarget* 2019;**10**:785–97.

Petukhov V, Xu RJ, Soldatov RA *et al.* Cell segmentation in imaging-based spatial transcriptomics. *Nat Biotechnol* 2022;**40**:345–54.

Quinlan JR. Induction of decision trees. *Machine Learning* 1986;**1**:81–106.

Rao A, Barkley D, França GS *et al.* Exploring tissue architecture using spatial transcriptomics. *Nature* 2021;**596**:211–20.

Rodriques SG, Stickels RR, Goeva A *et al.* Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science* 2019;**363**:1463–7.

Saha SK, Choi HY, Kim BW *et al.* KRT19 directly interacts with β-catenin/RAC1 complex to regulate NUMB-dependent NOTCH signaling pathway and breast cancer properties. *Oncogene* 2017;**36**:332–49.

Ståhl PL, Salmén F, Vickovic S *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* 2016;**353**:78–82.

Stickels RR, Murray E, Kumar P *et al.* Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nat Biotechnol* 2021;**39**:313–9.

Stringer C, Wang T, Michaelos M *et al.* Cellpose: a generalist algorithm for cellular segmentation. *Nat Methods* 2021;**18**:100–6.

Tan AC, Gilbert D. Ensemble machine learning on gene expression data for cancer classification. *Appl Bioinformatics* 2003;**2**:S75-83.

Vickovic S, Eraslan G, Salmén F *et al.* High-definition spatial transcriptomics for in situ tissue profiling. *Nat Methods* 2019;**16**:987–90.

Vizoso F, Plaza E, V□zquez J *et al.* Lysozyme expression by breast carcinomas, correlation with clinicopathologic parameters, and prognostic significance. *Ann Surg Oncol* 2001;**8**:667–74.

Wang J-J, Lei K-F, Han F. Tumor microenvironment: recent advances in various cancer treatments. *Eur Rev Med Pharmacol Sci* 2018;**22**:3855–64.

Wang X, Allen WE, Wright MA *et al.* Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* 2018;**361**:eaat5691.

Wang Y, Eddison M, Fleishman G *et al.* EASI-FISH for thick tissue defines lateral hypothalamus spatio-molecular organization. *Cell* 2021;**184**:6361-6377.e24.

Xia C, Fan J, Emanuel G *et al.* Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. *Proc Natl Acad Sci U S A* 2019;**116**:19490–9.

Zhang X, Ren D, Guo L *et al.* Thymosin beta 10 is a key regulator of tumorigenesis and metastasis and a novel serum marker in breast cancer. *Breast Cancer Res* 2017;**19**:15.