LOCATOR: feature extraction and spatial analysis of the cancer tissue microenvironment using mass cytometry imaging technologies

Rezvan Ehsani[1,2,*], Inge Jonassen[1,2] , Lars Akslen[2,3] and Dimitrios Kleftogiannis[1,2,*]

[1]Department of Informatics, Computational Biology Unit, University of Bergen, Bergen, Norway

[2]Centre for Cancer Biomarkers CCBIO, Department of Clinical Medicine, University of Bergen, Bergen, Norway

[3]Department of Pathology, Haukeland University Hospital, Bergen, Norway

[*]To whom correspondence should be addressed.

ABSTRACT

Motivation: Recent advances in highly multiplexed imaging have provided unprecedented insights into the complex cellular organization of tissues, with many applications in translational medicine. However, downstream analyses of multiplexed imaging data introduce several technical barriers, and although some computational methods and bioinformatics tools are available, deciphering the complex spatial organisation of cellular ecosystems remains a challenging problem.

Results: We developed a novel computational method, LOCATOR (ana**L**ysis **O**f **CA**ncer **Ti**ssue micr**O**envi**R**onment), for spatial analysis of cancer microenvironments using data acquired from mass cytometry imaging (MCI) technologies. LOCATOR introduces a graph-based representation of tissue images to describe features of the cellular organisation and deploys downstream analysis and visualisation utilities that can be used for data-driven patient risk stratification. Our case studies using MCI data from two well-annotated breast cancer cohorts re-confirmed that the spatial organisation of the tumour-immune microenvironment is strongly associated with the clinical outcome in breast cancer. In addition, we report interesting potential associations between the cellular organization of macrophages and patients' survival. Our work introduces an automated and versatile analysis framework for MCI data with many applications in future cancer research projects.

Availability: Datasets and codes of LOCATOR are publicly available at https://github.com/RezvanEhsani/LOCATOR.

Keywords

single-cell analysis, Imaging mass cytometry, systems biology, breast cancer, tumor microenvironment, bioinformatics, machine learning, graph analytics

INTRODUCTION

Mass cytometry imaging (MCI) is becoming an important tool for basic science and clinical research. So far, the prevalent technologies namely Imaging Mass Cytometry (IMC) and Multiplexed Ion Beam Imaging (MIBI), have found several healthcare applications including single cell studies in cancer, diabetes, and other complex immune diseases. With MCI clinical samples can be simultaneously analysed for up to 40 proteins/markers with spatial resolution, providing unprecedented insights into the complex multicellular organization of tissues. Especially in cancer research, several studies have shown that the heterogeneity of cell phenotypes, immune cell positioning, and cell-to-cell interactions within the tumour-immune microenvironment (TIME) are dominant factors of the clinical outcome. TIMEs have been also associated with responses to treatments including immunotherapies (Angelo, et al., 2014; Cabrita, et al., 2020; Eisenbarth, 2019; Giesen, et al., 2014; Groom, 2019; Helmink, et al., 2020; Lin, et al., 2018; Petitprez, et al., 2020) .

The growth of MCI technologies and the MCI data explosion has led to the development of computational methods for tissue exploration and data-driven patient stratification (Chen, et al., 2020; Ferro, et al., 2021; Goltsev, et al., 2018; Vickovic, et al., 2019). In a seminal study, Keren and colleagues used spatial analysis techniques to stratify a cohort of 41 triple-negative breast cancer (TNBC) patients and associated specific tissue organizations with immune regulation and patient survival (Keren, et al., 2018). Similarly, Stoltzfus and colleagues developed tools for spatial analyses of cells focusing on specific phenotypic neighbourhoods, that were used to explore the myeloid cell localization in lymph nodes of murine samples (Stoltzfus, et al., 2020). Other seminal studies focusing on spatial analyses of tissue microenvironments improved our overall understanding about phenotypic-genotypic associations with patient's outcome in breast cancer (Ali, et al., 2020; Jackson, et al., 2020). More recently, Daneberg et al. analysed large-scale (>690 samples) imaging data from breast tumours and demonstrated the importance of "connectivity features" within TIMEs to link pheno-genomic features with specific clinical outcomes (Danenberg, et al., 2022).

Although the existing methods opened new routes for deciphering spatial relationships between cell phenotypes and the clinical outcome, there are several technical barriers. A common limiting factor of MCI is that the resulting datasets are complex, spatial and multiparametric, thus challenging to integrate for downstream exploration. Typically, downstream analyses are not automated, relying on combinations of different software tools/packages that are often either proprietary and/or written in different programming languages. As a result, skilled personnel are required with prior knowledge of programming, system administration and data science. In addition, existing methods and tools are mainly designed to answer study-specific questions, and they cannot

be readily extended or generalized to investigate the role of arbitrary cell types of interest within TIMEs. Moreover, the lack of standardized approaches and mathematical models to describe cell-to-cell relationships limit our capacity to generalize findings and compare results from different studies that could be even acquired from different MCI technologies. Finally, there is lack of standard approaches to engineer numerical variables (i.e., features) of TIMEs, and use them to improve patient stratification via quantitative sciences and artificial intelligence (AI). Altogether, there is an increasing need for dedicated tools for hypothesis testing and data-driven MCI research to enhance our capacity for biomarker discoveries in the new era of precision medicine.

To mitigate these limitations, here we introduce LOCATOR (ana**L**ysis **O**f **CA**ncer **Ti**ssue micr**O**envi**R**onment) a novel spatial analysis method for MCI data from cancer patients. LOCATOR is implemented in R language that is widely used in biomedical research, and it is specifically designed to engineer features describing local/global cellular composition of tissues, as well as structural features of cells and their interactions within TIMEs. LOCATOR oversimplifies the analysis of MCI data allowing for powerful data-driven discoveries using graph-based feature engineering, machine learning and other knowledge discovery techniques. In addition, LOCATOR can be used for sample comparisons and facilitate easy and automated exploration of possible associations between TIMEs and other molecular and clinical variables. Importantly, LOCATOR is versatile, independent of the imaging technology used (e.g., IMC or MIBI), and can be applied to investigate any cell population of interest screened with any antibody panel.

To validate the potentials of the proposed method, we perform proof-of-concept studies using publicly available data from two breast cancer cohorts. Our results corroborate previous findings about the role of TIMEs and their association with patient survival. With LOCATOR we also investigate novel multicellular features of macrophages and report potentially interesting associations with distinct clinicopathological parameters. Our results convincingly demonstrate that LOCATOR is a general computational framework for spatial analyses of MCI data that could find many useful applications in future cancer research projects.

RESULTS

LOCATOR introduces a graph-based representation of cellular ecosystems for feature discoveries

The tumour microenvironment is composed of different cell types that form spatially organized cellular ecosystems. Within these cellular neighbourhoods immune and tumour cells interact directly or indirectly between each other adopting different mechanisms and functions. These functions play

important roles in disease progression, and many studies have emphasized on the role of TIMEs to suppress or promote tumour growth (Biswas, et al., 2022; Fu, et al., 2021; Hsieh, et al., 2022; Yuan, 2016). With this idea in mind, we rationalize that grouping together cellular neighbourhoods using salient spatial/cellular variables can be useful to understand the complex cancer tissue architecture, with the possibility to provide better patient stratification.

To contribute to this challenging task, we present LOCATOR. LOCATOR generates local neighbourhoods in the tissue microenvironments using a user-defined radius focusing on a pre-selected cell type of interest. To gain maximised information about each individual neighbourhood and subsequently generate comprehensive spatial features (i.e., numerical variables), LOCATOR adapts a graph-based representation of cellular ecosystems. With this abstract formulation all cells in the neighbourhood are modelled as vertices and links between them are represented as edges (i.e., cell-cell interactions). In this way, LOCATOR defines and computes three different categories of features namely: 1) cellular abundance (%) in the neighbourhood; 2) features describing the connectivity of cell types in the neighbourhood; 3) features described by standard graph theory metrics. A more detailed technical overview of the feature matrix generated by LOCATOR is described in the Materials and Methods section. Figure 1 shows a blueprint of LOCATOR's feature engineering modules. After the feature generation step, LOCATOR deploys an unsupervised consensus clustering approach to group together cellular neighbourhoods with similar/dissimilar properties as described by numerical features that are used as input. Patients can be enriched or depleted for each of the identified cellular clusters allowing to infer direct associations between TIME compositions and patient-specific clinicopathological parameters. The proposed data engineering approach is quite generic enabling effective downstream analyses, using for example artificial intelligence algorithms or any other data science technique.

LOCATOR systematises TIME-driven patient stratification

To showcase the use of LOCATOR for MCI data exploration and biomarker discoveries, the next subsections present results from two independent case studies in breast cancer focusing on Keren and Danenberg publicly available datasets (Danenberg, et al., 2022; Keren, et al., 2018). Choosing cell types of interest for LOCATOR analyses depends on the study design and the biological questions to be answered. Here, since one of the most promising clinical applications of MCI is to investigate the role of immune cells in the cellular microenvironment, our case studies are anchored around immune cell types, with a primary focus on macrophages.

Macrophages are innate immune cells with phagocytic function and heterogeneous characteristics, which play an important role in homeostasis and defence against pathogens (Qiu, et

al., 2018). In breast cancer high infiltration of tumour-associated macrophages (TAM) is associated with shorter patient survival (Choi, et al., 2018; Kumari and Choi, 2022; Lin, et al., 2019). However, our knowledge about the role and possible associations of macrophages with patient's survival is rather limited, and thus it is interesting to provide deeper insights into the spatial organization of macrophages within TIMEs. For extra experimentation with LOCATOR we also provide computational analysis results for another immune cell type, the T cytotoxic cells. The role of tumour-infiltrating lymphocytes (TILs) in cancer is well established (Dushyanthen, et al., 2015; Nascimento, et al., 2022; Stanton and Disis, 2016). Typically, high levels of TILs have shown significant prognostic and predictive value in specific breast cancer subgroups (HER2+, and Triple-negative). TIL profiling has indicated that presence of CD8+ T cytotoxic cells is strongly associated with improved prognosis, whilst more recent evidence reported their ability to block metastases of carcinogen-exposed breast cancers (Li, et al., 2021).

For validation purposes our results from TIME-specific patient stratification are associated with independent clinicopathological information (Supplementary Figure 1b). We also use published results obtained using completely independent computational approaches that are "blind" from LOCATOR's framework. In this way, although we provide indirect evidence, we use reasonable criteria highlighting LOCATOR's ability to test hypotheses and discover novel TIME-dependent patient subgroups that can be used for data-driven research.

Case study 1: LOCATOR application in Keren dataset

First, we apply LOCATOR to analyse 33 out of 41 TNBC samples available in Keren dataset. Five samples from this dataset are "cold" tumours (i.e., very low immune presence) and they are filtered out. We also filter out three samples with less than 20 macrophages to ensure robust feature matrices calculations for all samples included in the study. Note that this parameter can be customized by the users depending on the study design and the available cells of interest. The consensus clustering approach identifies four cellular clusters for macrophages (NC1, NC2, NC3 and NC4), and the mean abundance is used as a cut-off to identify enriched and depleted patients for each of these clusters. Figure 2 summarises LOCATOR's standard outputs that include a heatmap feature vector representation, summary plots focusing on the cellular neighbourhoods surrounding macrophages, a tsne visualisation of the clustering results and one example of tissue reconstruction mappings with clustering annotations. Similar outputs focusing on cellular neighbourhoods surrounding T cytotoxic cells are shown in Supplementary Figure 2.

Based on the consensus clustering results for macrophages we used LOCATOR to identify potential associations with disease outcome. To do so, cellular clustering results at the patient level

(i.e., number of patients enriched or depleted for each of NC1-NC4 clusters) are associated with the available clinicopathological parameters of the dataset. Since these parameters are independent from the data generation and clustering processes used by LOCATOR, they support unbiased evidence to assess the importance of our data-driven exploration. First, we use the available survival information to perform univariate Kaplan-Meier analysis. Figure 3a shows the survival analysis results for cluster NC2. We observe that patients enriched for this cluster (called NC2+) achieve worse survival compared to the depleted (NC2-) patients (long-rank p-value=0.018). Upon inspection of the connectivity features (Figure 2a) we observe that cluster NC2 has higher feature values for tumour-macrophages connectivity scores, whereas from the graph metric-based features the number of vertices is also high. Such findings might be linked to the underlying biology of TAMs within TIMEs, opening avenues for deciphering TAM's complex mechanisms.

Next, we use the mixing score published by Keren et al. to characterize the tissue architecture of breast cancers. For sanity, we compare our cellular clustering results with mixed and compartmentalized tissue types. Overall, we achieve good concordance between our clustering results and the independent results published by Keren et al. Figure 3b shows that patients of the NC2+ subgroup belong mainly to the mixed tissue type, which is linked with poorer survival, whereas patients of NC2- subgroup have mainly compartmentalized tissue types and better survival.

To support further our findings, we compare the inferred NC2 patient subgroups with the PDL1 immunoregulatory protein score levels. In the original study published by Keren et al. higher PDL1 score levels are associated with poorer survival, mixed tissue type and cancer progression. Our results (Figure 3c) focusing on NC2 cellular cluster of macrophages corroborate all the previous findings about PDL1 levels and tissue architecture. In summary, the results of this case study focusing on macrophages, agree with independent published results by Keren and colleagues, highlighting LOCATOR's ability for more generic data-driven research of TIMEs.

Case study 2: LOCATOR application in Danenberg dataset

We also apply LOCATOR to analyse 497 out of 693 samples from Danenberg dataset. We filter out 196 samples with less than 20 macrophages, 497 samples to ensure robust feature matrix calculation. Focusing on macrophages the consensus clustering approach identifies 14 cellular clusters (NC1 to NC14). We use the mean abundance as a cut-off to determine enriched and depleted patients for each of these cellular clusters. Figure 4 shows LOCATOR's outputs for this dataset including a heatmap feature vector representation, summary statistics focusing on the cellular neighbourhoods surrounding macrophages, a tsne visualisation of the clustering results and one example of tissue

reconstruction mappings with clustering annotations. Similar outputs focusing on cellular neighbourhoods surrounding T cytotoxic cells for this dataset are shown in Supplementary Figure 3.

Using LOCATOR, we investigate whether the acquired cellular clustering results are associated with patient survival. Based on univariate Kaplan-Meier analysis we find that patients enriched for NC2 (NC2+) and NC8 (NC8+) cellular clusters are linked with shorter survival whereas patients enriched for cellular cluster NC3 (NC3+) are associated with longer survival (Figure 5 a, b, and c).

To provide further evidence, we cross-check our cellular clustering results with annotations reported by Danenberg et al. using independent computational/experimental approaches (Figure 5 and Supplementary Figure 4). Focusing on NC2+ patient subgroup that achieve shorter survival, we find that this cellular cluster includes increased number of cases from IntClust 4- and IntClust 10 clusters that are both linked to poor prognosis. In NC2+ patient subgroup we also find proportionally less cases from IntClust 8 and IntClust 7 clusters that achieve good prognosis based on the published results. Focusing on PAM50 criteria, we find that NC2+ patient subgroup includes mainly Her2+ and Basal-like cases, whereas the number of LumA cases, which are in principle associated with favourable outcome, is proportionally smaller. Based on histological grade information, we observe that NC2+ subgroup includes more Grade 3 cases. Using ER hormone receptor status, we also see proportionally greater number of ER- cases for the NC2+ patient subgroup.

Focusing on NC3+ patient subgroup that achieve longer survival compared to NC3-, we find (Figure 5 and Supplementary Figure 4) that NC3+ includes proportionally more Normal Like cases. Based on histological grade information, we find that NC3+ patient subgroup includes proportionally smaller number of Grade 3 cases, and as expected contains mainly patients with Grade 1. Based on ER hormone receptor status, NC3+ patients include smaller number of ER- and more ER+ cases compared to NC3-.

Finally, for NC8+ patient subgroup that achieve shorter survival compared to NC8-, we observe that NC8+ includes proportionally decreased number of cases from IntClust 3, IntClust 7, and IntClust 8 which are associated with good survival; whereas it includes more cases from IntClust 5- cluster that is linked with poor survival. Focusing on PAM50 criteria, we find (Figure 5) that this patient subgroup includes proportionally smaller number of LumA cases and more cases with Her2. According to histological grade, NC8+ patient subgroup includes more Grade 3 cases (Supplementary Figure 4). At last, based on ER hormone receptor status, we observe that NC8+ patients are in their majority ER- cases. Supplementary Figure 5 shows additional clustering analysis results focusing on T cytotoxic cells. Here, we show three clusters (Supplementary Figure 3), namely NC2, NC3 and NC4, that are potentially linked with clinically relevant features of TIMEs. Taken together, the results of this case study using

LOCATOR's utilities to analyse the spatial neighbourhoods of macrophages and T cytotoxic cells, corroborate findings from independently published results by Danenberg and colleagues.

MATERIALS AND METHODS

LOCATOR overview

Required input data: LOCATOR consists of two programs namely *ExtractFeatures* and *TIMEClust* written in the R language. The method requires the availability of a set of patient samples screen with any of the available tissue imaging technologies using an antibody panel. Note that LOCATOR is a downstream analysis method, and thus it does not take as input "raw" images in tiff format. Instead, it takes as input single-cell data matrices in a tabular format (e.g., csv) where columns correspond to sample IDs, cell IDs, antibody markers, phenotype annotation (i.e., cell type), a binary variable indicating tumour or non-tumour cell type, followed by clinical annotations about the samples (e.g., PAM50 group, breast cancer grade or hormone receptor status). Rows in the data matrices correspond to individual cells. This tabular data representation is quite simple, and generic, and it can be obtained using any of the available MCI data pre-processing pipelines. For tissue image segmentation we recommend the state-of-the-art frameworks DeepCell (Bannon, et al., 2021), and ImcSegmentationPipeline (Vito RT Zanotelli, 2022). Once the segmented images have been produced, it is also required to identify phenotypes and annotate the available single cells using any of the available state-of-the-art approaches. Both unsupervised clustering methods and semi-automated supervised learning approaches are suitable for this task. Once the required input data have been obtained, the user must select the cells of interest. A standard approach is to select any of the "canonical" phenotypes (e.g., T-cells, B-cells, Macrophages etc) identified during the annotation pre-processing step. However, with LOCATOR it is also possible for users to select subsets of cells that are enriched/depleted based on the expression of an antibody marker (e.g., all immune cells enriched for PDL1).

Feature discovery: LOCATOR uses program *ExtractFeatures* to engineer features of TIMEs. LOCATOR deploys a versatile graph-based representation of cellular neighbourhoods to engineer features. To do so, for every cell of interest LOCATOR creates a circular neighbourhood using a predefined radius centred around every cell of interest. Choosing an appropriate value for the radius depends on the resolution of the imaging technology used. We recommend radius=50 pixels for MIBI and radius=100 for Hyperion IMC technologies. Within each neighbourhood LOCATOR constructs an individual indirected graph where cells are vertexes and interactions between them are edges. To identify all possible cell-to-cell interactions within a neighbourhood, the Voronoi algorithm is used (Aurenhammer, 1991). To speedup this process Voronoi algorithm considers only the cells within the neighbourhood, neglecting possible interactions of cells between different neighbourhoods (e.g.,

circles that are far apart). Based on this approach, three different categories of features are computed for every neighbourhood/graph (i.e., number of neighbourhoods equals the number of cells of interest) to capture phenotypic, functional, and structural information.

1) ***Cellular abundance (%)*** within the neighbourhood/graph: where the number of features equals the number of phenotypes found in the dataset.
2) ***Connectivity score*** **(CS)**: a weighted score reflecting the number of interactions and their importance within the neighbourhood. For each cell type within the neighbourhood, we first compute the number of interactions found with all other cell types (i.e., number of edges). Then each edge is weighted with *w1* if both vertexes (cells) are of the same cell type, with *w2* if the vertexes are of different cell types but they are both non-tumour cells or both tumour cells, and with *w3* if the vertexes are of different cell type but one is non-tumour cell and the other is tumour cell. Default weight parameters are *w1=1, w2=0.5 and w3=0.1*, but this can be adjusted by the user. The final connectivity score for each cell type within a neighbourhood, is computed by summing up all partial weighted scores.
3) ***Graph theory metrics***: for every neighbourhood/graph we compute the number of vertices, the number of edges, average vertex degree, entropy, density, diameter, eccentricity (max, min, mean), transitivity, algebraic connectivity, and spectral radius. Standard definitions of these metrics can be found in ref. Geng, Li 2012.

Note that the feature extraction process is performed once for each cohort and cell type of interest, and it can be used multiple times for subsequent downstream analysis tasks.

Data visualization and exploration: Once the feature matrix has been calculated, it can be fed to the TIMEClust program. TIMEClust deploys an unsupervised clustering approach to group together single cells based on their neighbourhoods' phenotypic, functional, and structural similarity. The consensus clustering approach is used from the Spectre package (Ashhurst, et al., 2022) to identify a statistically robust number of clusters, without manual tuning. Based on this unbiased stratification of cellular neighbourhoods LOCATOR offers several data exploration and visualization utilities. Users can generate and export boxplots and bar plots to visualize the abundance of clustered cells, and subsequently produce summary heatmaps of the feature matrices. LOCATOR integrates a dimensionality reduction approach using tsne, to project the cellular clusters into the 2D space. Single cell clusters can be also mapped to their original tissue coordinates providing as an output high-resolution tissue images for data inspection and exploration.

Furthermore, TIMEClust offers utilities for data-driven exploration at the cohort level. For each cellular cluster patients can be enriched or depleted given their underlying cellular abundance.

Supplementary Figure 6 shows examples of tissue images that are enriched and depleted for specific cellular clusters. At the patient level, this simple quantification can be used as a basis for data-driven patient stratification and association with clinical and biological annotations. LOCATOR integrates survival analysis based on Kaplan-Meier method, and survival plots are automatically generated for every cellular cluster. A list of Alluvial plots is also produced to visualize cellular clusters and their association with clinicopathological information. Command-line syntax for running *ExtractFeatures* and TIMEClust programs is provided on our github page. More detailed information about the dependent R packages and examples/tutorials are also available.

MCI cohorts and clinicopathological information used in this study

For the development and evaluation of LOCATOR we use publicly available MCI data from Keren et al. (Keren, et al., 2018) and Danenberg et al. (Danenberg, et al., 2022) herein we call them as "Keren dataset" and "Danenberg dataset" respectively. In the first dataset/study, Keren et al. used multiplexed ion beam imaging by time-of-flight (MIBI-TOF) to quantify 36 markers in 41 TNBC patients. The authors deployed an in-silico scoring scheme to investigate how immune and tumour cells interact with each other in respect to immune cell interactions. Using this scoring scheme, TIME architectures were associated with Mixed and Compartmentalized tissue types and validated that the Compartmentalized type was associated with improved survival. The authors also explored the expression of PDL1 immunoregulatory protein, in association to the spatial structure of the Compartmentalized and Mixed tissue types.

In the second dataset/study, Danenberg at el. used Hyperion IMC to generate 37-dimensional images of 693 breast cancer tumours matched with genomic and clinical information. Using spatial tissue-based analyses, the authors identified ten TIME structures/clusters, and explored how these clusters (IntClust information is available) were linked with patient-specific somatic alterations, ER status, tumour grade, and PAM50 breast cancer subgroups. Summary statistics about the datasets and patient-specific annotations used in both studies/datasets can be found in Supplementary Figure 1.

Discussion

In this study we present the development and use of LOCATOR, a new computational approach to analyse highly multiplexed imaging data from cancer tissues. LOCATOR integrates information on cellular phenotypes abundance, cellular connectivity, and cell to cell spatial relationships to shed light into the complex organisation of cellular microenvironments. LOCATOR deploys a versatile graph-based representation of cellular neighbourhoods to engineer features. Subsequently, it offers visualization utilities as well as a simple unsupervised clustering approach to stratify patients based on the similarity of their cellular neighbourhoods. LOCATOR programs are automated, they are

customizable by users, and they can be widely applied irrespective of the study design, the imaging technology used, and the available antibody panels. In this way, LOCATOR open possibilities to standardize MCI downstream analyses contributing towards hypothesis testing and exploration of cellular microenvironments via MCI.

To test LOCATOR, we perform two case studies using publicly available data from two seminal studies in breast cancer. Since our knowledge about the role of macrophages for patient stratification is limited, we use LOCATOR to analyse cellular microenvironments surrounding macrophages. Focusing on Keren dataset our patient stratification results, are in concordance with clinicopathological parameters and other computationally derived published findings. This is also the case for a larger cohort of MCI samples (Danenberg dataset), where LOCATOR contributes an alternative way for patient stratification based on the cellular organisation of macrophages. This expands possible applications of MCI to quantify tumour-immune infiltration using different immune cell types. Importantly, our findings are in agreement with previously published results and analyses focusing mainly on T cells. LOCATOR is in principle applicable to any other cell type of interest, and it runs fast (approximately 15 min for feature generation and downstream analyses for ~500 samples) in a commodity computer, without requiring significant programming knowledge. Thus, LOCATOR opens interesting possibilities to complement existing patient-risk stratification methods and enhance data-driven biomarker discoveries in future studies.

Focusing on technical aspects, many technical improvements are feasible. For example, implementing a graphical-user interface (GUI) would make LOCATOR even more interactive and easier for users with no prior programming knowledge skills. We are working towards this direction, providing improved customization of LOCATOR's utilities. In addition, incorporating state-of-the-art feature selection methods (Manikandan, 2021) to select the most important features of TIMEs, might be useful for more effective data-driven patient risk stratification tasks. In addition, other machine learning-based methods can be incorporated for biomarker discoveries (Tislevoll, et al., 2023). Furthermore, the proposed graph-based representation of cellular neighbourhoods can be subjected to more advanced graph algorithms and analytic platforms taking advantage of high-performance computing infrastructures. In terms of data analyses, the results presented here focusing mainly on macrophages corroborate many clinically relevant characteristics of patients. However, not all the detected patient clusters reach statistically significant associations with clinicopathological parameters. This is also the case for many of our clustering analysis results focusing on T-cytotoxic cells from Keren dataset. We argue that this is expected given the complexity of the underlying datasets and the performance of the unsupervised clustering approach we deployed. However, both case studies presented here highlight

LOCATOR's ability to test hypotheses and unbiasedly explore and stratify patients using features of their TIMEs'.

In summary, LOCATOR streamlines the application of tissue imaging technologies for data-driven patient stratification. The outcome of this work may be useful for cancer research studies focusing on cellular heterogeneity, and precision cancer medicine. LOCATOR's graph-based representation can be easily extended to describe even more comprehensive molecular profiles from single-cell multimodal assays and spatial transcriptomics (e.g., 10X Visium). In conclusion, we believe that the use of LOCATOR combined with artificial intelligence modelling, could provide a more general bioinformatics framework to explore the spatial organization of cellular communities in the broad context of human diseases.

Acknowledgements

Funding: This work it was supported by the Research Council of Norway through its Centre of Excellence funding scheme, project number 101651184. It was also partially supported by grants provided by EU ERA PerMed AML_PM (RCN Grant no. 298842).

Author Contribution

R.E developed LOCATOR and analysed the data; L.A and I.J provided intellectual input and help edit the manuscript; D.K and R.E wrote the manuscript; D.K. supervised the research.

Declaration of interests

The authors declare they have no competing interests.

References

Ali, H.R.*, et al.* Imaging mass cytometry and multiplatform genomics define the phenogenomic landscape of breast cancer. *Nat Cancer* 2020;1(2):163-175.

Angelo, M.*, et al.* Multiplexed ion beam imaging of human breast tumors. *Nat Med* 2014;20(4):436-442.

Ashhurst, T.M.*, et al.* Integration, exploration, and analysis of high-dimensional single-cell cytometry data using Spectre. *Cytometry A* 2022;101(3):237-253.

Aurenhammer, F. Voronoi Diagrams – A Survey of a Fundamental Geometric Data Structure. *ACM Computing Surveys* 1991;23(3):345–405.

Bannon, D.*, et al.* DeepCell Kiosk: scaling deep learning-enabled cellular image analysis with Kubernetes. *Nat Methods* 2021;18(1):43-45.

Biswas, A.*, et al.* Inference on spatial heterogeneity in tumor microenvironment using spatial transcriptomics data. *Comput Syst Oncol* 2022;2(3).

Cabrita, R.*, et al.* Author Correction: Tertiary lymphoid structures improve immunotherapy and survival in melanoma. *Nature* 2020;580(7801):E1.

Chen, Z.*, et al.* Modeling Multiplexed Images with Spatial-LDA Reveals Novel Tissue Microenvironments. *J Comput Biol* 2020;27(8):1204-1218.

Choi, J.*, et al.* The role of tumor-associated macrophage in breast cancer biology. *Histol Histopathol* 2018;33(2):133-145.

Danenberg, E.*, et al.* Breast tumor microenvironment structures are associated with genomic features and clinical outcome. *Nat Genet* 2022;54(5):660-669.

Dushyanthen, S.*, et al.* Relevance of tumor-infiltrating lymphocytes in breast cancer. *BMC Med* 2015;13:202.

Eisenbarth, S.C. Dendritic cell subsets in T cell programming: location dictates function. *Nat Rev Immunol* 2019;19(2):89-103.

Ferro, S.*, et al.* Phenomapping of Patients with Primary Breast Cancer Using Machine Learning-Based Unsupervised Cluster Analysis. *J Pers Med* 2021;11(4).

Fu, T.*, et al.* Spatial architecture of the immune microenvironment orchestrates tumor immunity and therapeutic response. *J Hematol Oncol* 2021;14(1):98.

Geng Li, M.S., Bülent Yener, Mohammed J. Zaki. Effective graph classification based on topological and label attributes. *Statistical analysis and data mining* 2012;5(4):265-283.

Giesen, C.*, et al.* Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat Methods* 2014;11(4):417-422.

Goltsev, Y.*, et al.* Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging. *Cell* 2018;174(4):968-981 e915.

Groom, J.R. Regulators of T-cell fate: Integration of cell migration, differentiation and function. *Immunol Rev* 2019;289(1):101-114.

Helmink, B.A.*, et al.* B cells and tertiary lymphoid structures promote immunotherapy response. *Nature* 2020;577(7791):549-555.

Hsieh, W.C.*, et al.* Spatial multi-omics analyses of the tumor immune microenvironment. *J Biomed Sci* 2022;29(1):96.

Jackson, H.W.*, et al.* The single-cell pathology landscape of breast cancer. *Nature* 2020;578(7796):615-620.

Keren, L.*, et al.* A Structured Tumor-Immune Microenvironment in Triple Negative Breast Cancer Revealed by Multiplexed Ion Beam Imaging. *Cell* 2018;174(6):1373-1387 e1319.

Kumari, N. and Choi, S.H. Tumor-associated macrophages in cancer: recent advancements in cancer nanoimmunotherapies. *J Exp Clin Cancer Res* 2022;41(1):68.

Li, K.*, et al.* CD8(+) T cell immunity blocks the metastasis of carcinogen-exposed breast cancer. *Sci Adv* 2021;7(25).

Lin, J.R.*, et al.* Highly multiplexed immunofluorescence imaging of human tissues and tumors using t-CyCIF and conventional optical microscopes. *Elife* 2018;7.

Lin, Y., Xu, J. and Lan, H. Tumor-associated macrophages in tumor metastasis: biological roles and clinical therapeutic applications. *J Hematol Oncol* 2019;12(1):76.
Manikandan, G., Abirami, S. Feature Selection Is Important: State-of-the-Art Methods and Application Domains of Feature Selection on High-Dimensional Data. *Applications in Ubiquitous Computing. EAI/Springer Innovations in Communication and Computing.* 2021.
Nascimento, C.*, et al.* The Landscape of Tumor-Infiltrating Immune Cells in Feline Mammary Carcinoma: Pathological and Clinical Implications. *Cells* 2022;11(16).
Petitprez, F.*, et al.* B cells are associated with survival and immunotherapy response in sarcoma. *Nature* 2020;577(7791):556-560.
Qiu, S.Q.*, et al.* Tumor-associated macrophages in breast cancer: Innocent bystander or important player? *Cancer Treat Rev* 2018;70:178-189.
Stanton, S.E. and Disis, M.L. Clinical significance of tumor-infiltrating lymphocytes in breast cancer. *J Immunother Cancer* 2016;4:59.
Stoltzfus, C.R.*, et al.* CytoMAP: A Spatial Analysis Toolbox Reveals Features of Myeloid Cell Organization in Lymphoid Tissues. *Cell Rep* 2020;31(3):107523.
Tislevoll, B.S.*, et al.* Early response evaluation by single cell signaling profiling in acute myeloid leukemia. *Nat Commun* 2023;14(1):115.
Vickovic, S.*, et al.* High-definition spatial transcriptomics for in situ tissue profiling. *Nat Methods* 2019;16(10):987-990.
Vito RT Zanotelli, B.B. ImcSegmentationPipeline: A pixel-classification based multiplexed image segmentation pipeline. *Zenodo* 2022.
Yuan, Y. Spatial Heterogeneity in the Tumor Microenvironment. *Cold Spring Harb Perspect Med* 2016;6(8).

Figure 1. LOCATOR workflow. The flowchart describes the data acquisition and pre-processing process of multiplexed imaging data that is required to generate the LOCATOR's input data in csv format. LOCATOR workflow uses two programs written in R language namely, ExtractFeatures and TIMEClust. ExtractFeatures is used to define cellular neighborhood and generate graphs. These graphs are used for feature extraction. Next, TIMEClust program uses these features to cluster the cellular neigbourhoods of interest. At the patient level, patients as enriched or decreased for each of the identified clusters. This provides opportunities for MCI data visualisation and downstream analyses for data-driven patient stratification tasks.

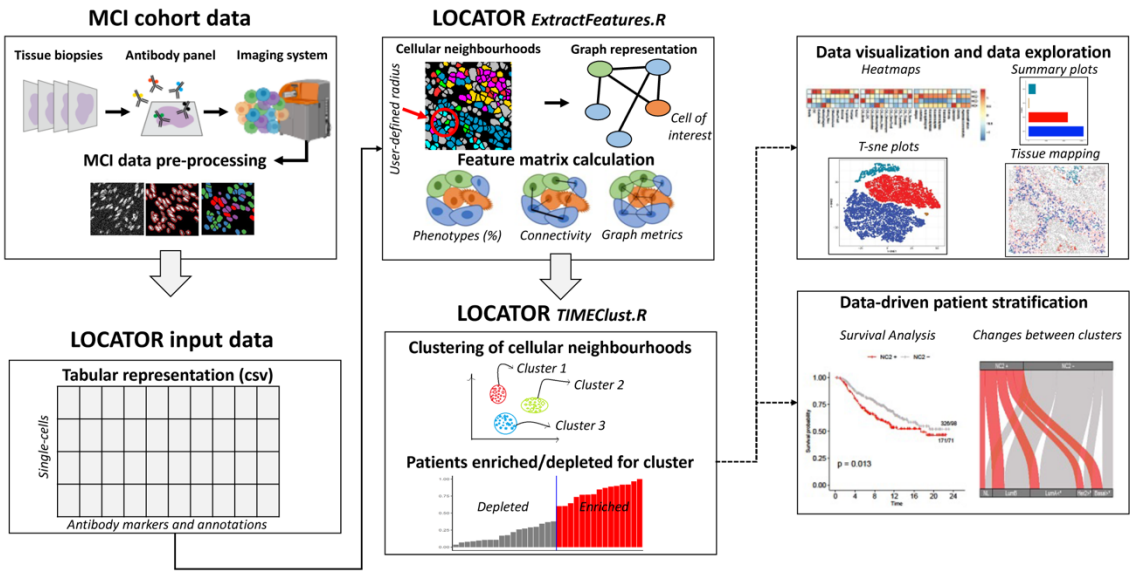


Figure 2. LOCATOR outputs for Keren dataset focusing on macrophage cell type. a) Heatmap representation of the feature vector for the identified clusters. b) Bar plot showing cellular abundance of the identified cellular clusters. c) Boxplot showing the number of macrophages per patients across different cellular clusters. d)T-sne visualisation of the identified cellular clusters. e) Example of tissue mapping with cellular cluster annotation from one patient (patient No 5). Gray colour indicates tumour cells and pink colour indicates non-tumour cells. All other colours in the image correspond to the detected cellular clusters from subplots c and d.

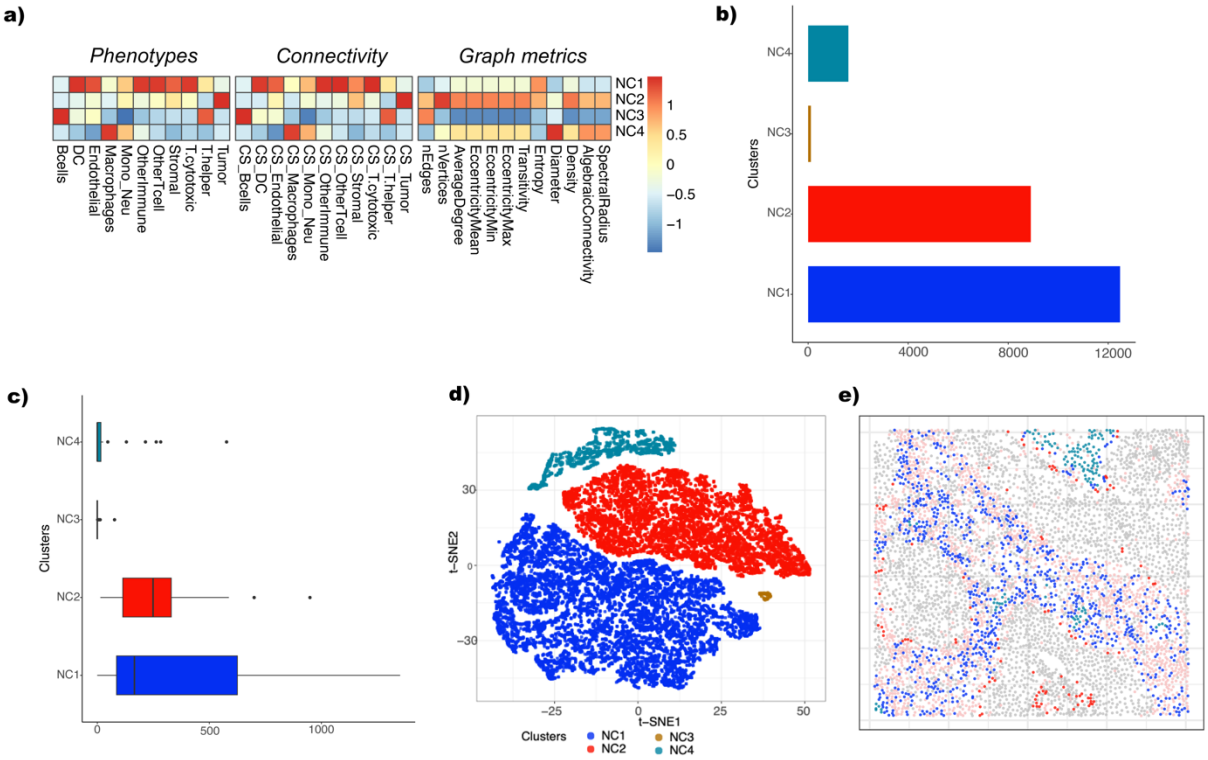


Figure 3. Case study 1: Association of the identified NC2 cellular cluster with clinicopathological parameters and independent published results. a) Survival plot where patients are classified as enriched (NC2+) or depleted (NC2-) based on their mean abundance for NC2. b) Alluvial plot showing the fraction of NC2 enriched/depleted patients and their underlying tissue mixing classification. c) From left to right, original classification of patients and PDL1 score from Keren et al., where colour code is based on the patients enriched/decreased for NC2 cellular cluster. The stack bar shows the abundance of different macrophage cellular clusters identified by TIMEClust across all patients.

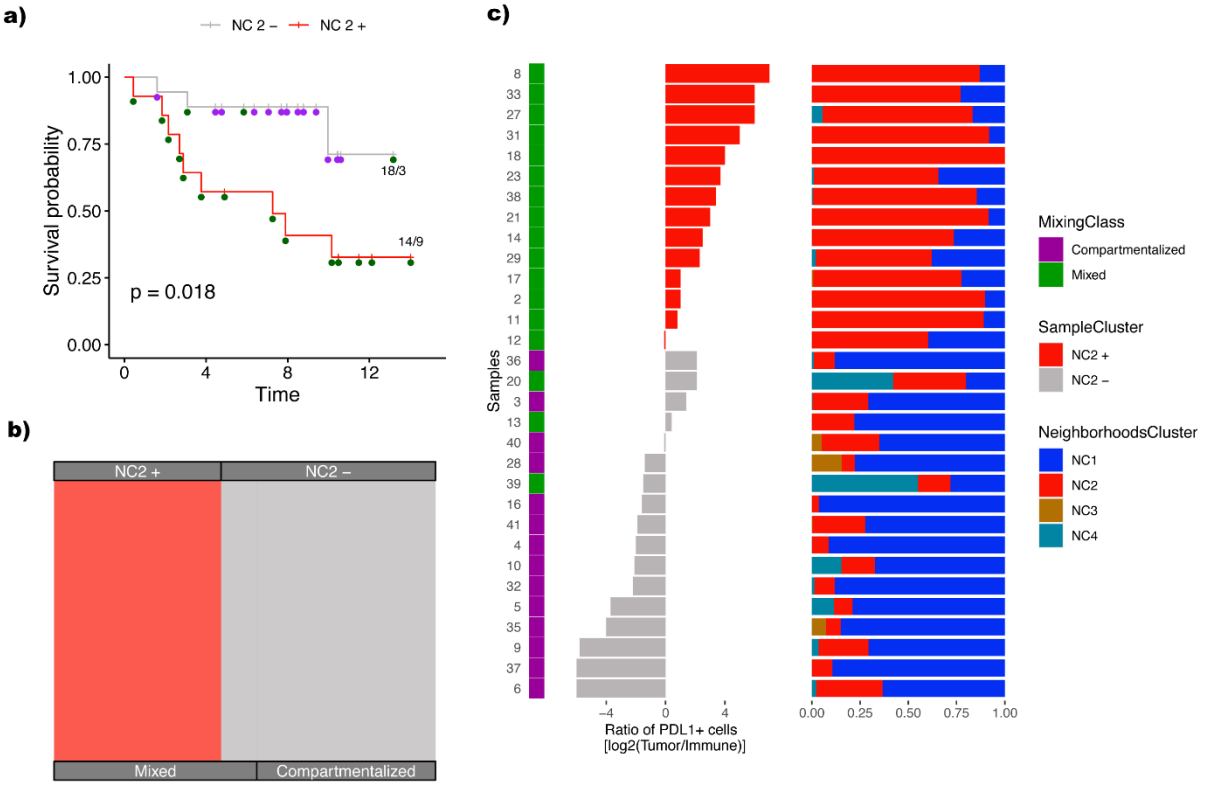


Figure 4. LOCATOR outputs for Danenberg dataset focusing on macrophage cell type. a) Heatmap representation of the feature vector for the identified clusters. b) Bar plot showing cellular abundance of the identified cellular clusters. c) Boxplot showing the number of macrophages per patients across different cellular clusters. d)T-sne visualisation of the identified cellular clusters. e) Example of tissue mapping with cellular cluster annotation from one patient (patient No 191). Gray colour indicates tumour cells and pink colour indicates non-tumour cells. All other colours in the image correspond to the detected cellular clusters from subplots c and d.

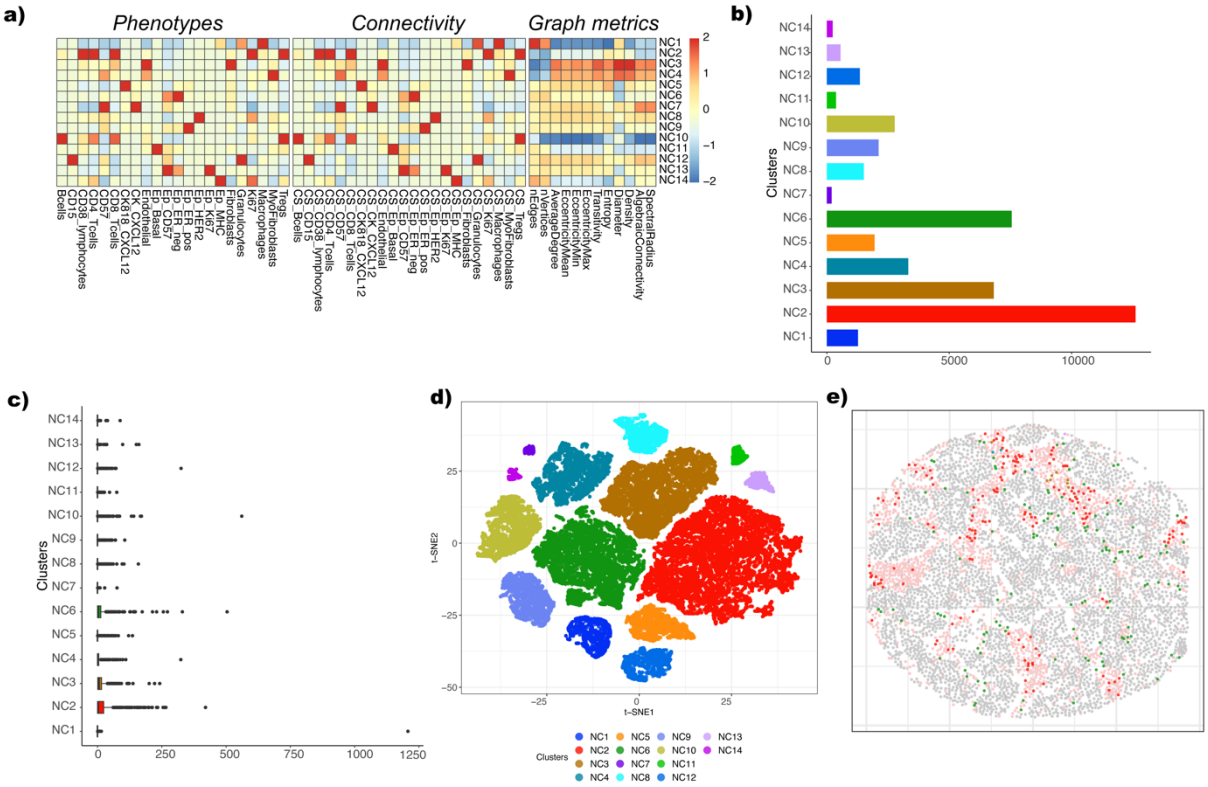
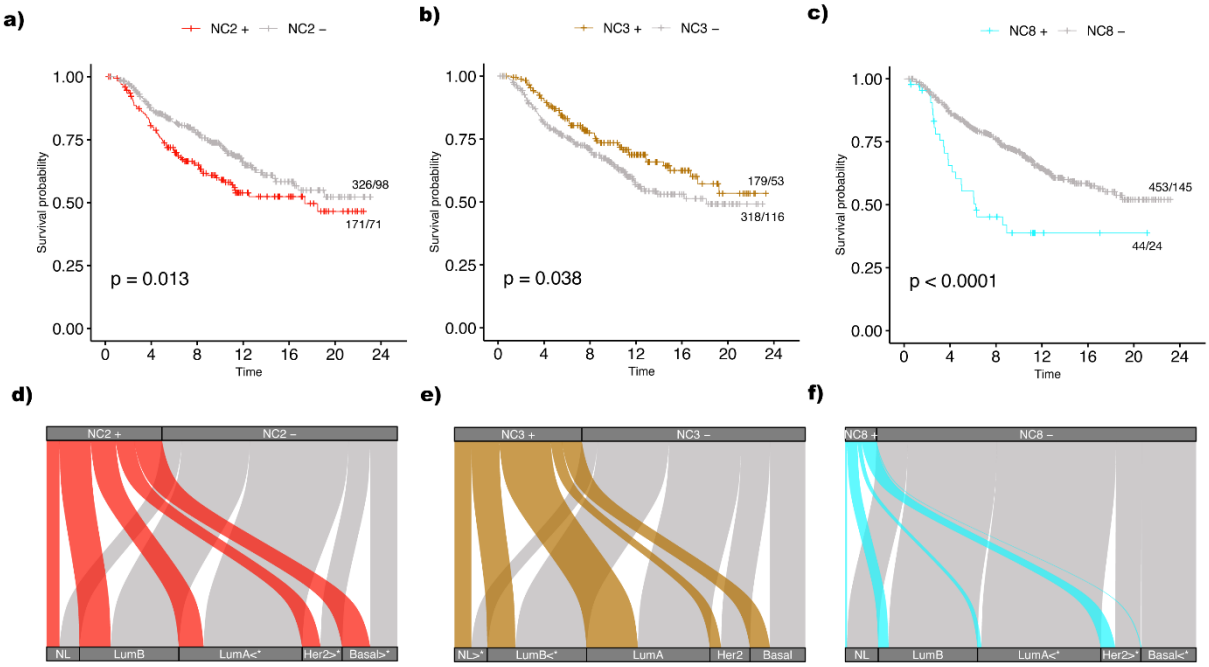


Figure 5. Case study 2: Association of NC2, NC3 and NC8 identified cellular clusters with patient survival and PAM50 classification. a-c) Survival plots where patients are classified as enriched (NC2+, NC3+, NC8+) or depleted (NC2- , NC3-, NC8-) based on their mean cellular abundances. d-f) Alluvial plots showing the fraction of NC2, NC3 and NC8 enriched/depleted patients and their underlying PAM50 classification. Significance of the results is obtained using Fishers-exact test at a p-value levels of 0.05 (results marked with *, and Supplementary Figure 7 for details).



Supplementary Figure 1. Summary statistics and annotations for Keren and Danenberg datasets.

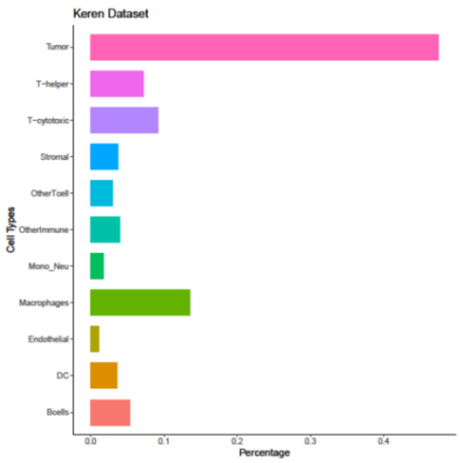
a)

Dataset	#sample	#markers	#cell	#cell types	#samples used in LOCATOR	Clinical Annotations
Keren et al.	41(36*)	36	190,953	11	33	Mixing score Survival
Danenberg et al.	693	37	1,045,835	22	497	PAM50 IntClust Tumour Grade ERStatus Survival

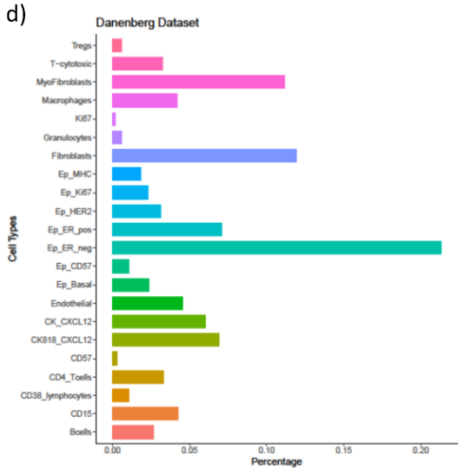
b)

	Annotations	Distinguishing Features	Prognosis
PAM50	NL (Normal Like)	ER+, PR+, HER2-	Good
	LumA (Luminal A)	ER+, PR+\PR-, HER2-	Good
	LumB (Luminal B)	ER+, PR+\PR-, HER2+\ HER2-	Intermediate
	Her2	ER-, PR-, HER2+	Poor
	Basal	ER-, PR-, HER2-	Poor
IntClust	IntClust 1	17q23 amplification GATA3 mutations High genomic instability	Intermediate
	IntClust 2	11q13-14 amplification (CCND1) High genomic instability	Poor
	IntClust 3	Low genomic instability with few copy number changes PIK3CA and CDH1 mutations	Good
	IntClust 4-	ER-negative tumors irrespective of HER2 status	Poor
	IntClust 4+	Low genomic instability Up regulation immune response genes	Good
	IntClust 5-	ERBB2 amplification	Poor
	IntClust 5+	ERBB2 abbreviation	Good
	IntClust 6	8p12 amplification (ZNF703) High genomic instability	Intermediate
	IntClus 7	16p gain, 16q loss, 8q amplification MAP3K1 mutations	Good
	IntClust 8	1q gain, 16q loss PIK3CA and GATA3 mutations	Good
	IntClust 9	8q gain, 20q amplification High genomic instability TP53 mutations	Intermediate
	IntClust 10	5q loss, 8q gain, 10p gain, 12p gain Impaired DNA checkpoint regulation, TP53 mutations	Poor
Grade	G1 (Grade 1)	slower-growing and less likely to spread	Good
	G2 (Grade 2)	growing faster than a grade 1 cancer but slower than a grade 3 cancer	Intermediate
	G3 (Grade 3)	faster-growing cancer that's more likely to spread	Poor
ER status	ER +	Cancer cells have receptors that allow them to use the hormone estrogen to grow	Good

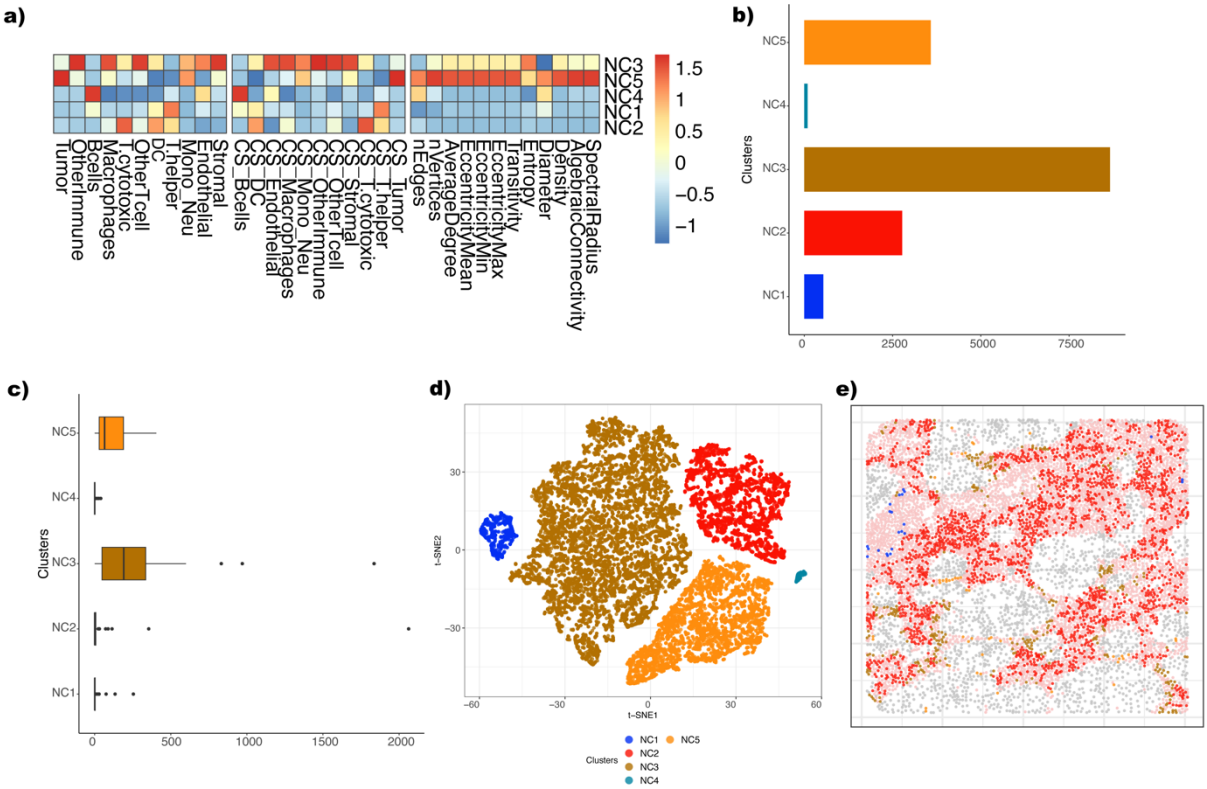
c)



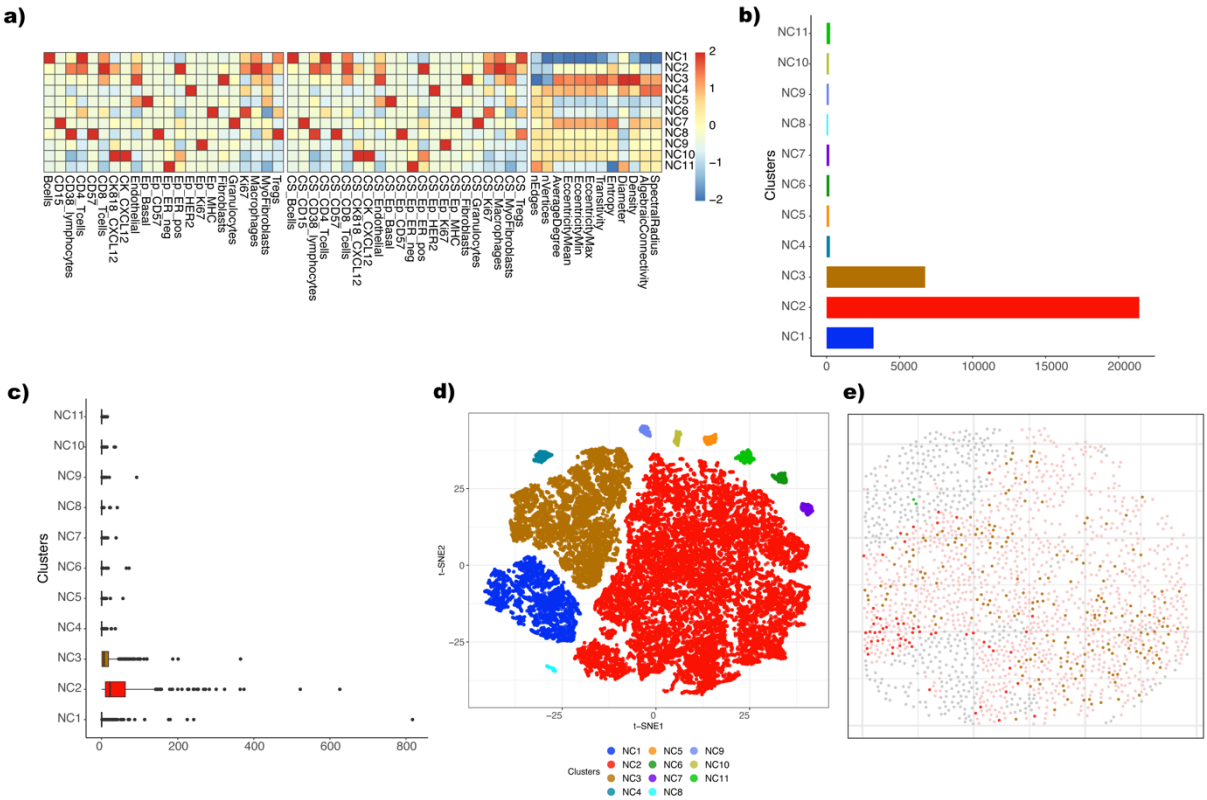
d)



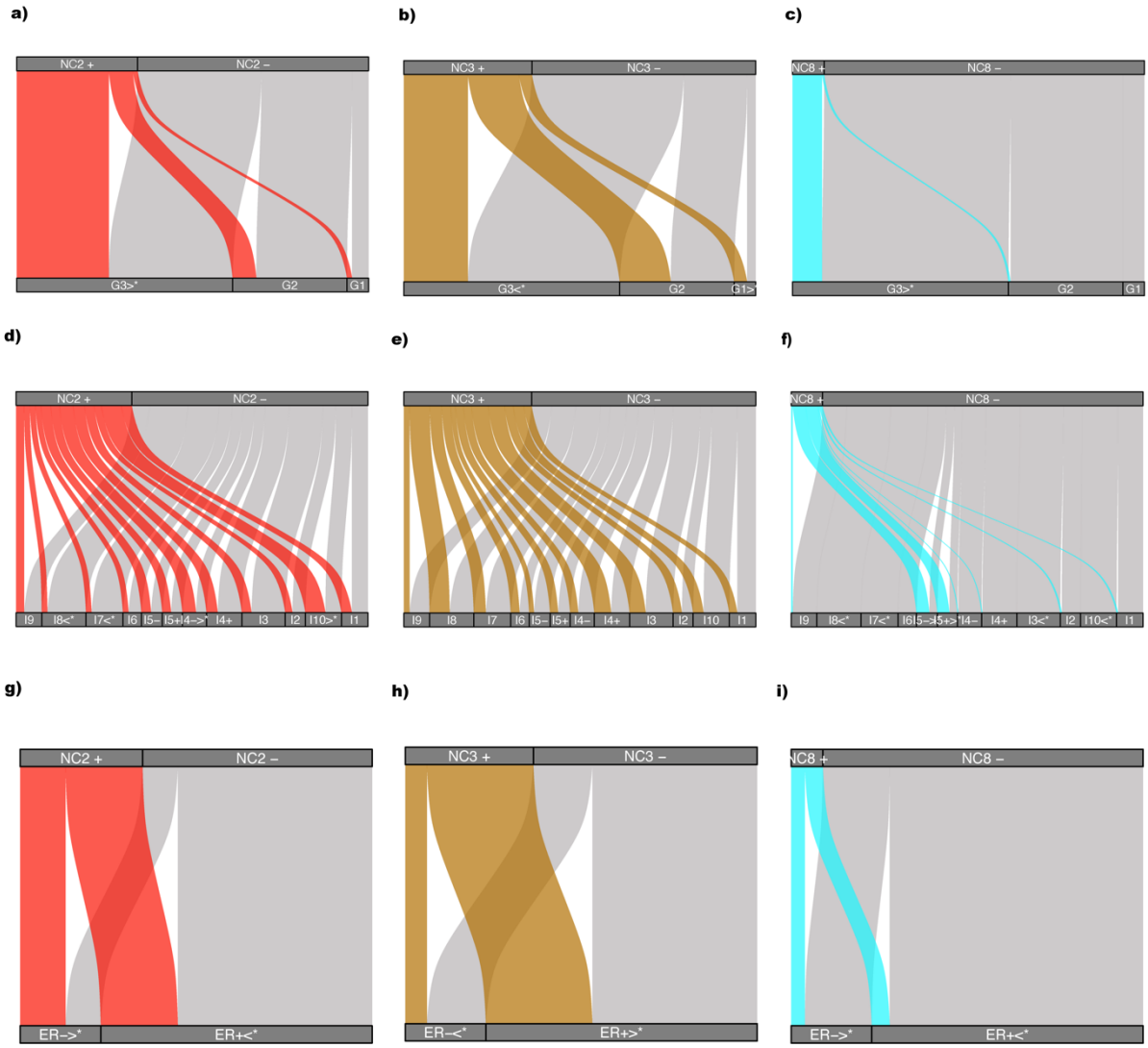
Supplementary Figure 2. LOCATOR outputs for Keren dataset focusing on T cytotoxic cells. a) Heatmap representation of the feature vector for the identified clusters. b) Bar plot showing cellular abundance of the identified cellular clusters. c) Boxplot showing the number of T-cytotoxic cells per patients across different cellular clusters. d)T-sne visualisation of the identified cellular clusters. e) Example of tissue mapping with cellular cluster annotation from one patient (patient No 7). Gray colour indicates tumour cells and pink colour indicates non-tumour cells. All other colours in the image correspond to the detected cellular clusters from subplots c and d.



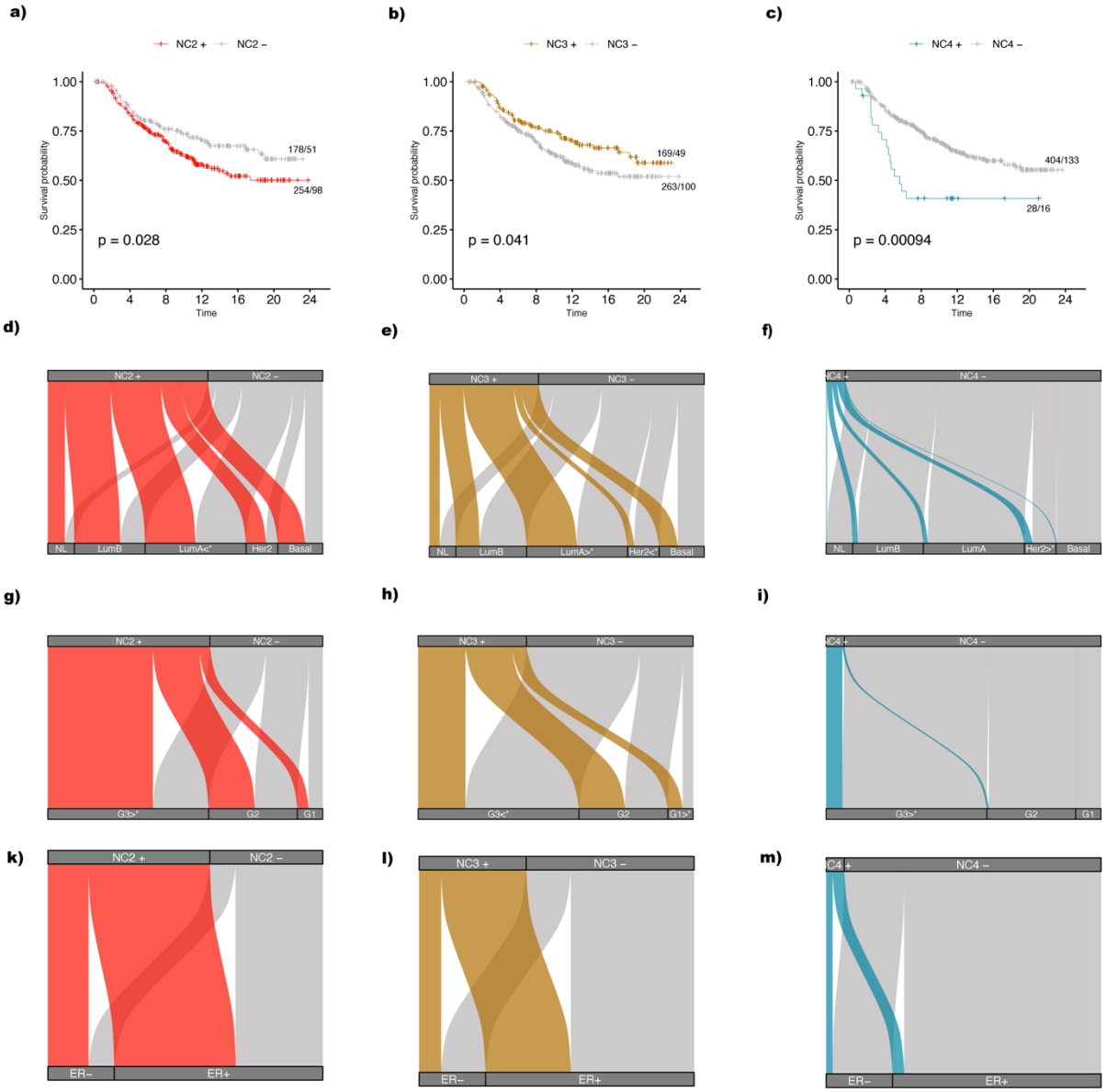
Supplementary Figure 3. LOCATOR outputs for Danenberg dataset focusing on T cytotoxic cells. a) Heatmap representation of the feature vector for the identified clusters. b) Bar plot showing cellular abundance of the identified cellular clusters. c) Boxplot showing the number of T-cytotoxic cells per patients across different cellular clusters. d)T-sne visualisation of the identified cellular clusters. e) Example of tissue mapping with cellular cluster annotation from one patient (patient No 47). Gray colour indicates tumour cells and pink colour indicates non-tumour cells. All other colours in the image correspond to the detected cellular clusters from subplots c and d.



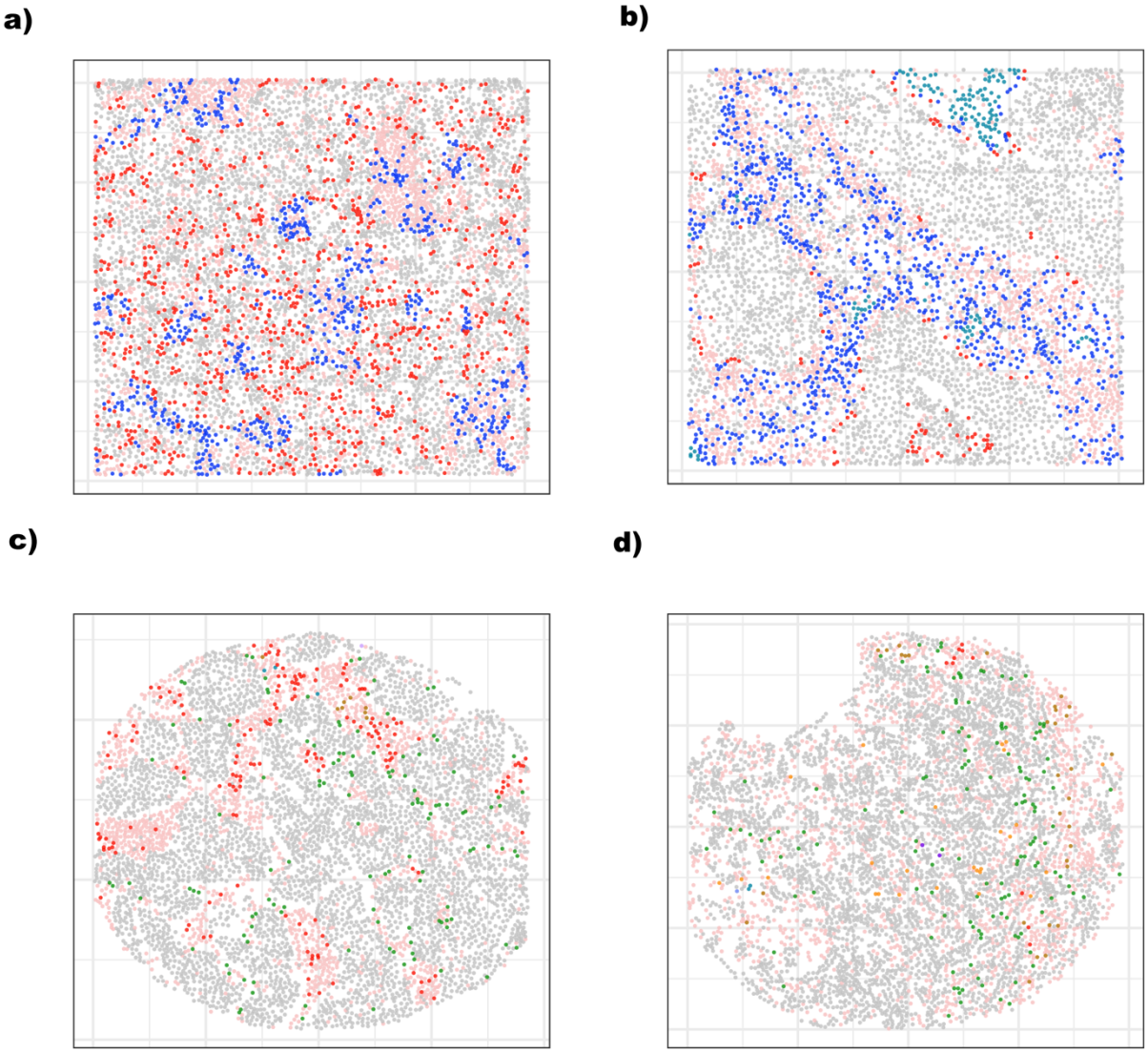
Supplementary Figure 4. Case study 2: Association of NC2, NC3 and NC8 identified cellular cluster with clinicopathological parameters and independent published results. a-i) Alluvial plots showing the fraction of NC2, NC3 and NC8 enriched/depleted patients and their underlying associations with histological Grade, IntClust results and ER status. Significance of the results is obtained using Fishers-exact test at a p-value level of 0.05 (results marked with *, and and Supplementary Figure 7).



Supplementary Figure 5. LOCATOR application on T-cytotoxic cells from Case study 2: Association of NC2, NC3 and NC4 identified cellular clusters with survival and other clinicopathological parameters. a-c) Survival plots where patients are classified as enriched (NC2+, NC3+, NC4+) or depleted (NC2- , NC3-, NC4-) based on their mean cellular abundances. d-m) Alluvial plots showing the fraction of NC2, NC3 and NC4 enriched/depleted patients and their underlying PAM50 classification, histological grade and ER status. Significance of the results is obtained using Fishers-exact test at a p-value levels of 0.05 (results marked with *, and Supplementary Figure 7).



Supplementary Figure 6. Examples of tissue reconstruction and mapping generated by LOCATOR visualisation utilities. a-b) Selected tissue images from Keren dataset. c-d) Selected tissue images from Danenberg dataset. In all subplots, fray colour indicates tumour cells and pink colour indicates non-tumour cells. All other colours correspond to cellular cluster annotations reported in Figure 2 and Figure 4.



Supplementary Figure 7: Summary results of the statistical tests (Fisher's exact test) we performed to associate cellular clustering findings with clinicopathological parameters.

CluterType	ClinicalType	+/-	+/+	-/-	-/+	P value
NC2	G3	0.24	0.76	0.46	0.54	0
NC2	ER+	0.37	0.63	0.15	0.85	0
NC2	ER-	0.63	0.37	0.85	0.15	0
NC2	Basal	0.76	0.24	0.88	0.12	0.001
NC2	LumA	0.78	0.22	0.58	0.42	0
NC2	Her2	0.84	0.16	0.91	0.09	0.039
NC2	IntClust 7	0.95	0.05	0.87	0.13	0.007
NC2	IntClust 8	0.95	0.05	0.84	0.16	0
NC2	IntClust 10	0.83	0.17	0.93	0.07	0.001
NC2	IntClust 4-	0.88	0.12	0.96	0.04	0.008
NC3	G3	0.5	0.5	0.32	0.68	0
NC3	G1	0.9	0.1	0.96	0.04	0.01
NC3	ER+	0.17	0.83	0.26	0.74	0.019
NC3	ER-	0.83	0.17	0.74	0.26	0.019
NC3	NL	0.87	0.13	0.93	0.07	0.042
NC3	LumB	0.78	0.22	0.68	0.32	0.028
NC8	G3	0.07	0.93	0.42	0.58	0
NC8	ER+	0.43	0.57	0.21	0.79	0.002
NC8	ER-	0.57	0.43	0.79	0.21	0.002
NC8	Basal	0.97	0.03	0.83	0.17	0.019
NC8	LumA	0.87	0.13	0.63	0.37	0.001
NC8	Her2	0.54	0.46	0.92	0.08	0
NC8	IntClust 7	1	0	0.88	0.12	0.024
NC8	IntClust 8	1	0	0.86	0.14	0.009
NC8	IntClust 5+	0.62	0.38	0.97	0.03	0
NC8	IntClust 3	1	0	0.86	0.14	0.009
NC8	IntClust 10	1	0	0.89	0.11	0.023
NC8	IntClust 5-	0.59	0.41	0.97	0.03	0